American Psychological Association

Volume 109 July 2017
Number 5

Published eight times

ISSN 0022-0663

Journal of Educational Psychology

www.apa.org/pubs/journals/edu

The main purpose of the *Journol of Educotional Psychology*® is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journol* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Single Issues, Back Issues, and Back Volumes: For information regarding single issues, back issues, or back volumes, write to Order Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242; call 202-336-5600 or 800-374-2721; or visit www.apa.org/pubs/journals/subscriptions.aspx

Manuscripts: Submit manuscripts electronically through the Manuscript Submissions Portal found at www.apa.org/pubs/journals/edu according to the Instructions to Authors found elsewhere in this issue (see table of contents). Correspondence regarding manuscripts should be sent to the Editor, Steve Graham, at steve.graham@asu.edu. The opinions and statements published are the responsibility of the authors, and such opinions and statements do not necessarily represent the policies of APA or the views of the Editor.

Journal of Educational Psychology® (ISSN 0022-0663) is published eight times (January, February, April, May, July, August, October, November) in one volume per year by the American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Subscriptions are available on a calendar year basis only (January through December). The 2017 rates follow: *Nonmember Individual:* $250 Domestic, $292 Foreign, $314 Air Mail. *Institutional:* $953 Domestic, $1,030 Foreign, $1,054 Air Mail. *APA Member:* $123. *APA Student Affiliate:* $75. Write to Subscriptions Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Printed in the U.S.A. Periodicals postage paid at Washington, DC, and at additional mailing offices. POSTMASTER: Send address changes to *Journal of Educational Psychology,* American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

Effective with the 1986 volume, this journal is printed on acid-free paper.

Journal of

Educational Psychology®

www.apa.org/pubs/journals/edu

July 2017

Volume 109
Number 5

Motivation

Learning

Mathematics

Reading

(*Contents continue*)

Other

Journal of Educational Psychology
2017, Vol. 109, No. 5, 599–620

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000153

How True Is Grit? Assessing Its Relations to High School and College Students' Personality Characteristics, Self-Regulation, Engagement, and Achievement

Katherine Muenks, Allan Wigfield, Ji Seung Yang, and Colleen R. O'Neal
University of Maryland

Duckworth, Peterson, Matthews, and Kelly (2007) defined *grit* as one's passion and perseverance toward long-term goals. They proposed that it consists of 2 components: consistency of interests and perseverance of effort. In a high school and college student sample, we used a multidimensional item response theory approach to examine (a) the factor structure of grit, and (b) grit's relations to and overlap with conceptually and operationally similar constructs in the personality, self-regulation, and engagement literatures, including self-control, conscientiousness, cognitive self-regulation, effort regulation, behavioral engagement, and behavioral disaffection. A series of multiple regression analyses with factor scores was used to examine (c) grit's prediction of end-of-semester course grades. Findings indicated that grit's factor structure differed to some degree across high school and college students. Students' grit overlapped empirically with their concurrently reported self-control, self-regulation, and engagement. Students' perseverance of effort (but not their consistency of interests) predicted their later grades, although other self-regulation and engagement variables were stronger predictors of students' grades than was grit.

Keywords: grit, personality, self-regulation, engagement

Supplemental materials: http://dx.doi.org/10.1037/edu0000153.supp

Grit, defined by Duckworth, Peterson, Matthews, & Kelly, 2007, as "trait-level perseverance and passion for long-term goals," encompasses one's ability to maintain interests, exert effort, and persist at tasks over long periods of time (p. 1087). Duckworth and her colleagues originally conceptualized grit within personality theory (e.g., John & Srivastava, 1999), describing grit as a trait that is similar to conscientiousness or self-control, but that relates specifically to long-term stamina toward goals (Duckworth et al., 2007). Over the past several years, policymakers, practitioners, and the general public have become interested in grit, and journalists have written popular press books and articles about the importance of grit in individuals' lives (Ivcevic & Brackett, 2014; Sehgal, 2015; Tough, 2012; U.S. Department of Education, 2013). However, to date there has not been much rigorous, empirical research on grit in different groups of people. Specifically, the measure of grit developed by Duckworth et al. has not been fully validated. Further, few researchers have examined how grit relates to conceptually and operationally similar constructs in the self-regulation and engagement literatures, as well as school-based achievement outcomes. Thus the goals of the present study were to (a) examine the factor structure of the most commonly used grit measure in high school and college samples; (b) assess the overlap and distinctiveness of grit with constructs from the personality, self-regulation, and engagement literatures; and (c) Examine the relations between grit and later achievement outcomes, specifically students' grades, when other variables are controlled.

Defining and Measuring Grit: Current Evidence on Its Factor Structure

Duckworth et al. (2007) conceptualized grit as having two components, consistency of interests and perseverance of effort. They created a scale to measure grit called the Grit-Original (or Grit-O) that included 12 items; six items tapped consistency of interests (e.g., "I often set a goal but later choose to pursue a different one") and six items tapped perseverance of effort ("I am diligent"; see Table 1 for all items). They defined perseverance of effort as individuals' tendencies to keep working toward long-term

This article was published Online First December 5, 2016.

Katherine Muenks, Allan Wigfield, and Ji Seung Yang, Department of Human Development and Quantitative Methodology, University of Maryland; Colleen R. O'Neal, Department of Counseling, Higher Education, and Special Education, University of Maryland.

The writing of this article was supported in part by a dissertation fellowship awarded to Katherine Muenks from the National Academy of Education/Spencer Foundation. Part of this work is supported by the National Science Foundation under Grant No. 1534846. Any opinions, findings, and conclusions or recommendations expressed in this manuscript are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Correspondence concerning this article should be addressed to Katherine Muenks, who is now at the Department of Psychological and Brain Sciences, Indiana University, 1101 East 10th Street, Room 389, Bloomington, IN 47405. E-mail: kmuenks@iu.edu

Table 1
All Items for all Scales

Construct	Item
Grit: Consistency of interests (Grit-CI; Duckworth & Quinn, 2009)	I often set a goal but later choose to pursue a different one. (R)
	I have been obsessed with a certain idea or project for a short time but later lost interest. (R)
	I have difficulty maintaining my focus on projects that take more than a few months to complete. (R)
	New ideas and projects sometimes distract me from previous ones. (R)
Grit: Perseverance of effort (Grit-PE; Duckworth & Quinn, 2009)	I finish whatever I begin.
	Setbacks don't discourage me.
	I am diligent.
	I am a hard worker.
Conscientiousness (Gosling et al., 2003)	I see myself as dependable, self-disciplined.
	I see myself as disorganized, careless. (R)
Self-control (Tangney et al., 2004)	I am good at resisting temptation.
	I have a hard time breaking bad habits. (R)
	I am lazy. (R)
	I say inappropriate things. (R)
	I do certain things that are bad for me, if they are fun. (R)
	I refuse things that are bad for me.
	I wish I had more self-discipline. (R)
	People would say that I have iron self-discipline.
	Pleasure and fun sometimes keep me from getting work done. (R)
	I have trouble concentrating. (R)
	I am able to work effectively toward long-term goals.
	Sometimes I can't stop myself from doing something, even if I know it is wrong. (R)
	I often act without thinking through all the alternatives. (R)
Cognitive self-regulation (Pintrich et al., 1991)	During class I often miss important points because I'm thinking of other things. (R)
	When reading for this course, I make up questions to help focus my reading.*
	When I become confused about something I'm reading for this class, I go back and try to figure it out.
	If course materials are difficult to understand, I change the way I read the material.
	Before I study new course material thoroughly, I often skim it to see how it is organized.
	I ask myself questions to make sure I understand the material I have been studying in this class.
	I try to change the way I study in order to fit the course requirements and instructor's teaching style.
	I often find that I have been reading for class but don't know what it was all about. (R)*
	I try to think through a topic and decide what I am supposed to learn from it rather than just reading it over when studying.
	When studying for this course I try to determine which concepts I don't understand well.
	When I study for this class, I set goals for myself in order to direct my activities in each study period.
	If I get confused taking notes in class, I make sure I sort it out afterwards.
Effort regulation (Pintrich et al., 1991)	I often feel so lazy or bored when I study for this class that I quit before I finish what I planned to do. (R)
	I work hard to do well in this class even if I don't like what we are doing.
	When course work is difficult, I give up or only study the easy parts. (R)
	Even when course materials are dull and uninteresting, I manage to keep working until I finish.
Behavioral engagement (Skinner et al., 2008)	I try hard to do well in school.
	In class, I work as hard as I can.
	When I'm in class, I participate in class discussions.
	I pay attention in class.
	When I'm in class, I listen very carefully.
Behavioral disaffection (Skinner et al., 2008)	When I'm in class, I just act like I'm working.
	I don't try very hard at school.
	In class, I do just enough to get by.
	When I'm in class, I think about other things.
	When I'm in class, my mind wanders.

Note. R = Reverse coded. Items with asterisks were given only to the college sample.

goals. This construct is similar to *future time perspective*, defined by Lens (1986) and Husman and Lens (1999) as "the degree to which and the way in which the chronological future is integrated into the present life-space of an individual through motivational goal-setting processes" (Husman & Lens, 1999, p. 114). However, because the items on the perseverance of effort subscale of the Grit-S (e.g., "I finish whatever I begin," "Setbacks don't discourage me," "I am diligent," "I am a hard worker") do not necessarily reflect long-term goals, the conceptualization of the construct may not be accurately reflected in the measure. We return to this issue later.

Duckworth and colleagues' (2007) consistency of interest scale captures their notion of passion: Individuals' tendencies to pursue the same or similar activities over time. It is important to note that Duckworth and colleagues' (2007) conceptualization of "consistency of interests" differs from the way interest is conceptualized by prominent researchers who study interest, such as Hidi and Renninger (2006). Hidi and Renninger (2006) distinguish "individual interest," which they define as "a relatively enduring predisposition to reengage particular contents over time" from "situational interest," which is triggered by current environmental stimuli (p. 111). Thus, Duckworth and colleagues' definition of consistency of interests is more goal- and action-oriented and encompasses long-term behavior, rather than reflecting a personal disposition toward a particular topic (individual interest) or interest that is triggered by a particular situation (situational interest). The broad point here, however, is that each proposed component of grit overlaps to a degree with constructs already in the literature.

Duckworth et al. (2007) conducted exploratory and confirmatory factor analyses on the Grit-O with samples of adults and confirmed that the two subscales, consistency of interests and perseverance of effort, were separate but correlated factors. Duckworth and Quinn (2009) then created a shorter, eight-item version of the grit scale, the Short Grit Scale (or Grit-S), by eliminating items from the Grit-O that were not as predictive of various outcomes across several samples of participants. These samples included West Point Academy students, other college students, and National Spelling Bee contestants. They tested a model positing the two subscales as first-order latent factors and grit as a second-order latent factor (e.g., a higher order latent factor model), and found that this model fit adequately for some samples but not well for others. These analyses provided some initial information about the hypothesized factor structure of the Grit-S, which is currently the most commonly used grit measure.

One important methodological issue that also has theoretical ramifications is whether the two subscales of grit proposed by Duckworth and colleagues (2007) are part of a single latent construct called *grit* or form two distinct latent constructs. In other words, is grit a single, cohesive construct with two subscales, or are the two subscales different enough to be different constructs? Duckworth and Quinn (2009) and Wolters and Hussain (2015) both tested a higher order factor model of grit, in which consistency of interests and perseverance of effort were posited as first-order factors, and grit a second-order factor. This would seemingly test the hypothesis that grit is a single construct with two subscales. However, because there are only two subscales, a higher order factor model of grit is mathematically equivalent to a two correlated-factor model, in which the two subscales are posed as separate, but correlated factors (e.g., Schmid & Leiman, 1957, and see Appendix C in the online supplemental materials). Thus, a higher order factor model is not the best model to test whether grit's two subscales make up a single, cohesive construct. Instead, a bifactor model of grit, in which one underlying factor is posited, and two subscales capture the residual dependencies among the items (e.g., Holzinger & Swineford, 1937), is a more appropriate model to test whether grit is a single construct.

Thus to address our first research goal, we tested three competing measurement models of grit: a one-factor model in which grit is a single construct with no subscales, a two correlated-factor model in which consistency of interests and perseverance of effort are separate but correlated constructs, and a bifactor model. We did not expect that the one-factor model would fit well, as Wolters and Hussain (2015) already demonstrated. However, we did not have any specific predictions about which of the other two models would fit best. If the bifactor model fits best, it would support Duckworth et al.'s (2007) conceptualization and operational definitions of grit as a single construct with two subscales. If the two correlated-factor model fits best, it would be unclear whether grit is a single construct or if the subscales are distinct constructs.

Another question that researchers have not yet systematically examined is whether grit's factor structure is consistent across different age groups, such as high school and college students. Researchers have examined grit in samples ranging in age from eight to adulthood (e.g., Duckworth & Quinn, 2009; Rojas & Usher, 2013; West et al., 2016; Wolters & Hussain, 2015); thus it is important to examine whether the factor structure may be different at various developmental levels. Based on previous work there are at least two possibilities for how grit's factor structure varies developmentally. The first is that the two subscales of grit, consistency of interests and perseverance of effort, become more differentiated as students get older and are better able to cognitively differentiate between the two, as has been found in work on other self-perception variables (Harter, 2006). Alternatively, the two proposed subscales of grit might become more highly correlated as students get older. As students transition from high school to college and begin to encounter more difficult material and set longer-term goals, they may also understand that accomplishing their goals requires both hard work and consistent interest over time. To address these possibilities, we compared the factor structures of grit in our high school and college samples.

Grit's "Nomological Network": Constructs Conceptually and Empirically Overlapping With Grit in Different Theoretical Frameworks

As mentioned earlier, grit emerged from personality theory, so Duckworth and colleagues (Duckworth et al., 2007; Duckworth & Quinn, 2009) have focused primarily on its conceptual and empirical relations to other personality traits. However, theorists in educational and developmental psychology proposing social–cognitive models of self-regulation and engagement include in their models constructs that are conceptually and operationally similar to grit (Pintrich & Schrauben, 1992; Skinner, Kindermann, Connell, & Wellborn, 2009; Zimmerman, 2011). To date there is not abundant work on the relations of these various constructs to grit; we next review the extant work. We acknowledge that this is not an exhaustive list and that there are constructs within theories of motivation, such as interest theory and achievement goal theory (e.g., Hidi & Renninger, 2006; Wolters, 2004) that are conceptually similar to aspects of grit. However, given that we did not measure these constructs in the present study, we do not review the literature on them.

Conceptually Similar Constructs to Grit From Personality Theory

Conscientiousness and grit. Conscientiousness is one of the Big Five personality traits and is defined as being hardworking, responsible, self-disciplined, and thorough (John & Srivastava, 1999). Duckworth et al. (2007) conceptually differentiated grit from conscientiousness by stating that grit refers to stamina and consistency of interests over long periods of time, something not mentioned in definitions of conscientiousness. Duckworth and colleagues (Duckworth et al., 2007; Duckworth & Quinn, 2009; Eskreis-Winkler, Shulman, Beal, & Duckworth, 2014) found that grit relates strongly to conscientiousness. To date, researchers have not examined whether items used to measure grit and conscientiousness would factor separately; that is, whether grit and conscientiousness would emerge as empirically distinct constructs.

Self-control and grit. Self-control, although not one of the Big Five personality traits, has been studied often in the personality literature (e.g., Mischel, Cantor, & Feldman, 1996). Various researchers (e.g., Mischel et al., 1996; Tangney, Baumeister, & Boone, 2004) define *self-control* as one's capacity to change oneself in order to create an optimal fit between the self and the world; a particularly important part of self-control is the ability to override impulses for undesired behaviors. Duckworth and Gross (2014) conceptually differentiated self-control from grit by stating that self-control refers to resisting short-term temptations, whereas grit refers to passion and effort sustained over the long term. However, as can be seen in Table 1 that presents all the measures, none of the items on the Grit-S refer to long-term goals, so it seems likely there would be much empirical overlap between these constructs. As is the case with grit and conscientiousness, researchers have not yet examined whether grit and self-control are empirically distinct by including items measuring both constructs in the same factor analysis.

Conceptually Similar Constructs to Grit From Social–Cognitive Theories of Self-Regulation and Engagement

Social–cognitive models of self-regulation and engagement, such as Zimmerman's (2011) influential model of self-regulation and Skinner's (e.g., Skinner et al., 2009) model of engagement both contain constructs that are conceptually quite similar to grit, and researchers have found that some of these variables correlate with grit (e.g., Christensen & Knezek, 2014; Rojas et al., 2012; Wolters & Hussain, 2015). We focus here on cognitive self-regulation and effort regulation from models of self-regulation, and behavioral engagement and behavioral disaffection from Skinner et al.'s (2009) engagement model.

Cognitive self-regulation and grit. Cognitive self-regulation includes cognitive learning strategies, such as planning, monitoring, regulating, and appraising before, during, and after performance (Zimmerman, 2011). Much research has shown that when students self-regulate their learning they perform better, maintain their interest in what they are doing, and have positive affect about what they are doing (see Pintrich & De Groot, 1990; Zimmerman, 2011). In these and other ways self-regulation overlaps conceptually with aspects of grit, such as regulating one's effort to attain long-term goals, maintaining interest, and having positive affect. A few studies have found that grit is related to aspects of self-regulation or self-regulated learning (Christensen & Knezek, 2014; Rojas et al., 2012; Wolters & Hussain, 2015).

Effort regulation and grit. Pintrich, Smith, Garcia, and McKeachie (1991) defined effort regulation as an aspect of self-regulation that refers to students' ability to maintain effort and attention to tasks, even when they become uninteresting or there are other distractions in the environment (Pintrich et al., 1991). Pintrich and de Groot (1990) found that effort regulation was associated with middle school students' intrinsic value, self-efficacy, and strategy use, as well as achievement. Conceptually, effort regulation seems quite similar to the perseverance of effort aspect of grit; however, no one has examined this empirically.

Behavioral engagement, behavioral disaffection, and grit. Engagement is a multidimensional construct that is associated with being deeply involved in an activity (Christenson, Reschly, & Wylie, 2012; Fredricks et al., 2004; Skinner et al., 2009; Skinner, Furrer, Marchand, & Kindermann, 2008; Skinner, Kindermann, & Furrer, 2008). Behavioral engagement encompasses effort exertion, persistence, attention, and concentration (Skinner et al., 2009), whereas behavioral disaffection involves passivity and wanting to opt out of achievement activities (Skinner et al., 2008). Behavioral engagement and disaffection seem particularly similar to the perseverance of effort component of grit. Researchers have also not examined whether grit and the different aspects of engagement are empirically distinct, or how strongly they relate to one another. To address the second goal of this study, we examined (using confirmatory item factor analyses) whether grit factored separately from or with each of the above constructs.

Grit and Achievement Outcomes

With respect to grit's relations to outcomes, researchers have found that individuals' grit predicts outcomes such as completion of training courses in the military and performance at the National Spelling Bee (e.g., Duckworth et al., 2007; Duckworth, Kirby, Tsukayama, Berstein, & Ericsson, 2011; Duckworth & Quinn, 2009; Eskreis-Winkler et al., 2014; Maddi, Matthews, Kelly, Villarreal, & White, 2012). Relations between grit and later school outcomes have also been found, such as students' achievement test score gains from fourth to eighth grade (West et al., 2016), graduation from high school (Eskreis-Winkler et al., 2014), grades in elementary and middle school (Rojas & Usher, 2013), college grades (Chang, 2014; Duckworth et al., 2007; Strayhorn, 2014), doctoral program grades (Cross, 2014), and years of education completed by adults (Duckworth et al., 2007; Duckworth & Quinn, 2009). In this work previous achievement was not typically controlled.

However, other researchers found no relations of grit to later academic outcomes, or that these relations disappear once researchers control for other variables. For example, Chang (2014) found that grit did not predict college students' grade point averages (GPAs), Cross (2014) found no relations of grit to the completion of the dissertation in doctoral students, and West et al. (2016) found no relations between eighth grade students' grit and achievement test scores. Finally, Ivcevic and Brackett (2014) reported that grit did not predict high school students' concurrent academic recognitions, honors, or GPA when other personality traits were controlled in the analyses. Thus, the strength of the

relation between grit and academic outcomes is inconsistent across studies.

To address our third research goal, we examined how grit predicted students' grades in a particular course when controlling for other personality, self-regulation, and engagement constructs. We chose grades as our outcome over standardized test scores because grades are arguably the achievement outcome that students care most about, they are predictive of students' later school and occupational opportunities and success (e.g., Geiser & Santelices, 2007; Hoffman & Lowitzki, 2005; Roth, BeVier, Switzer, & Schippmann, 1996; Roth & Clarke, 1998; Thorsen & Cliffordson, 2012), and they represent students' achievement over the course of several months.

Purposes of the Present Study

To address our research goals in this study, we addressed three research questions:

1. What is the best-fitting factor structure model of grit for high school and college students—a one-factor model, a two correlated-factor model, or a bifactor model?

2. How empirically distinct or overlapping are grit and the conceptually similar constructs of conscientiousness, self-control, cognitive self-regulation, effort regulation, behavioral engagement, and behavioral disaffection? As discussed in more detail in the analysis section, we did these analyses pair by pair, in order to obtain the greatest clarity with respect to their overlap.

3a. Does students' grit predict their later grades after controlling for gender and ethnicity?

3b. Which constructs (grit, conscientiousness, self-control, cognitive self-regulation, effort regulation, behavioral engagement, or behavioral disaffection) are the most powerful independent predictors of grades after controlling for gender and ethnicity?

3c. Does students' grit predict their later grades after controlling for gender, ethnicity, and similar constructs?

Method

Sample 1 (High School)

Participants. Participants were 203 high school juniors (50.7% female) attending a private high school in the mid-Atlantic region of the United States. Participants were recruited via their classrooms through a partnership between researchers and administrators at the school. We obtained passive consent from parents, and active assent from students. The average age was 16.33 years (SD = .51 years). The ethnic breakdown was 61.6% White, 12.6% Black, 11.1% Biracial, 7.6% Latino, 4.5% Asian/Asian American, and 2.5% other.

Measures.

Grit. Grit was measured with the Grit-S (Duckworth & Quinn, 2009). The scale consists of eight items. Previous studies found adequate internal consistency (αs = .70 to .84; Duckworth & Quinn, 2009). Four items tap students' consistency of interests (e.g., "I often set a goal but later choose to pursue a different one"), and four items tap students' perseverance of effort (e.g., "I am diligent"). Participants indicated their agreement with each item on a scale from 1 = *not at all like me* to 5 = *very much like me.*

Conscientiousness. Conscientiousness was measured with the Conscientiousness subscale of the Ten-Item Personality Inventory (TIPI; Gosling, Rentfrow, & Swann, 2003), which consists of two items. The TIPI has been used in many previous studies (e.g., Back et al., 2010; Carney, Jost, Gosling, & Potter, 2008; Correa, Hinsley, & Zuniga, 2010). Participants were asked to indicate the extent they see themselves as "dependable, self-disciplined" and "disorganized, careless" (reverse-coded) on a scale from 1 = *disagree strongly* to 7 = *agree strongly.* These two items correlated at $r = -.42$ in a previous study (Gosling et al., 2003).

Self-control. Self-control was measured with the Brief Self-Control Scale (Tangney et al., 2004). The scale consists of 13 items (e.g., "I am good at resisting temptation"). Participants indicated their agreement with each item on a scale from 1 = *not at all like me* to 5 = *very much like me.* In previous studies this scale had good internal consistency (αs = .83 to .85; Tangney et al., 2004).

Cognitive self-regulation. Cognitive self-regulation was measured with the metacognitive self-regulation scale from the Motivation Strategies for Learning Questionnaire (MSLQ; Pintrich et al., 1991). The original scale consists of 12 items measuring one's planning, monitoring, and regulating activities during learning (e.g., "If course materials are difficult to understand, I change the way I read the material"); however, two items ("When reading for this course, I make up questions to help focus my reading" and "I often find that I have been reading for class but don't know what it was all about") were removed from the scale for the high school sample because they were not as relevant to high school students (e.g., there are not always "readings" for high school classes like there are in college classes, especially for math and science classes). Thus, students responded to 10 items on a scale from 1 = *not at all true of me* to 7 = *very true of me.* In a previous study, the internal consistency for the full scale was adequate (α = .79; Pintrich et al., 1991). When responding to the items, participants were asked to think about a math or science class that they were currently enrolled in.[1]

Effort regulation. Effort regulation was measured with the effort regulation scale from the MSLQ (Pintrich et al., 1991). The scale consists of four items (e.g., "I work hard to do well in this class even if I don't like what we are doing"). Students responded on a scale from 1 = *not at all true of me* to 7 = *very true of me.* In previous research, the internal consistency was adequate (α = .69; Pintrich et al., 1991). When responding to the items, participants were asked to think about a math or science class that they were currently enrolled in.

Behavioral engagement and behavioral disaffection. Behavioral engagement and disaffection with learning were measured with the Engagement versus Disaffection with Learning

[1] We asked students to think about a math or science class because we thought that grit might be particularly important for these subjects, which students typically view as more challenging than others (e.g., Linn & Eylon, 2006; Maloney, 2016).

Scale (Skinner et al., 2008). The scale consists of subscales for behavioral engagement (sample item: "In class, I work as hard as I can") and behavioral disaffection (sample item: "I don't try very hard at school"). In a previous study, the internal consistencies for behavioral engagement (α = .61) and behavioral disaffection (α = .71) were adequate (Skinner et al., 2008). Participants indicated their agreement with each item on a scale from 1 = *not at all true* to 4 = *very true*.

Grades. After the semester ended, students' grades were collected via school records for the class they thought about when responding to the self-regulation and effort regulation scales. Because each semester consisted of two quarters, teachers calculated students' end-of-semester grades using their final grades from the first two quarters (40% for each quarter) and the final exam (20%) in the course. These three grade variables had good internal consistency (α = .82). Grades were in the form of percentages and were on a scale from 1 to 100.

Procedure. Participants completed a series of questionnaires during the school day in January (end of fall semester) of their junior year. Assent forms, instructions, and the questionnaires themselves were all provided and completed digitally on students' personal iPads, which every student at the school had. After completing the assent form, participants were given (in a randomized order) scales measuring grit, self-control, engagement/disaffection with learning, and conscientiousness. Items within the scales were randomized. Next, because the self-regulation and effort regulation scales were designed to be taken with reference to a particular course, participants were asked to think about a specific class in which they were currently enrolled, and to type the name of the course they would be thinking about when responding to the self-regulation and effort regulation scales. Then, they were given these scales in a randomized order.[2] Finally, participants were asked to report some additional information about the course and their educational background, as well as their demographics. After the semester ended, participants' grades in that class and their overall GPAs were collected via school records.

Sample 2 (College)

Participants. Participants were 336 undergraduate students (74.4% female) from a mid-Atlantic university. The average age was 20.16 years (*SD* = 2.65 years). The ethnic breakdown was 58.1% White, 18.5% Asian/Asian American, 9.4% Black, 7.3% Latino, 6.1% Biracial, and 0.6% other.

Seventy-three professors who were teaching courses during the fall semester were contacted by the first author and asked if they would be interested in helping to recruit participants for a research study. They were asked if they would be willing to offer extra credit to their students for participation, or, if not, whether they would be willing to send a link to the survey to their students for voluntary participation. Six professors agreed to give extra credit and 17 professors agreed to send the link for voluntary participation. The large majority of participants came from the courses where the professors offered extra credit; only 28 participants were volunteers. Students came from a variety of different courses: 54.4% from the behavioral and social sciences; 24.2% from agriculture/natural resources; 12.4% from computer, mathematical, and natural sciences; and 9% from courses in a field not specified by the survey.

Measures. The scales (and items) were almost identical to those given to the high school sample, except the college students completed the full cognitive self-regulation scale, noted below. After providing consent, participants were given (in a randomized order) measures of grit, self-control, engagement/disaffection with learning, and conscientiousness. Next, participants were asked to think about the specific course they were recruited from, since the self-regulation and effort regulation scales were designed to be taken with reference to a particular course. They were asked to indicate which course they would be thinking about and then completed the self-regulation and effort regulation scales in a randomized order. Items within all scales were randomized. Finally, they were asked to report some additional information about the course, their educational background, and their demographics.[3]

Cognitive self-regulation. Cognitive self-regulation was measured with the full metacognitive self-regulation scale from the MSLQ (Pintrich et al., 1991) that consisted of 12 items.

Grades. After the semester ended, students' grades were collected via school records for the class they thought about when responding to the self-regulation and effort regulation scales. The grades were in the form of letters (e.g., A–, B+, etc.) and were recoded on a scale from 1 to 13 where 1 = F and 13 = A+.

Procedure. All professors sent a link to their students through email and the students completed the survey at home on their own time during the middle of the fall semester. At that time students also gave permission for researchers to access their final grades in the course. After the semester was complete, students' grades were collected via the registrar's office.

See Tables 2 and 3 for means and standard deviations for all variables in the high school and college samples. The descriptive statistics are very similar across high school and college samples. It should be noted that the means and standard deviations are based on the summed scores that assume interval scales for the Likert item responses. After testing the assumption, further analyses were conducted treating all the Likert item responses as categorical variables.

Analysis plan. Missing data for the high school sample ranged from 1.5% to 4.9%, and for the college sample ranged from 0.9% to 3.6%, for individual items used in the analyses. For the multidimensional item response theory (MIRT) analyses (explained in more detail subsequently), we used the full information maximum likelihood estimator, which can be seen as a model-based approach to missing data (see Enders, 2010). For the regression analyses (described subsequently), we used pairwise deletion given the minimal (less than 5%) missing data rates (see Schafer, 1999).

Before describing the analyses used to address the research questions, we discuss some initial decisions concerning whether to treat the item response data as continuous or categorical and which estimation techniques to use in the item factor analyses.

Pre-data analytic decisions.

Likert scaled response data: Continuous versus categorical. An important consideration when analyzing data from Likert

[2] Participants in both studies (high school and college) also responded to a few additional questionnaires, but analyses pertaining to those measures are not reported here.

[3] None of the analyses using this information are reported here.

Table 2
Correlations, Means, Standard Deviations, and Reliability Coefficients for all Variables—High School Sample

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Grit-CI													
2. Grit-PE	.55**												
3. Conscientiousness	.40**	.64**											
4. Self-control	.58**	.52**	.52**										
5. Cognitive self-reg	.33**	.47**	.32**	.43**									
6. Effort reg	.38**	.57**	.46**	.48**	.78**								
7. Behavioral eng	.30**	.57**	.43**	.48**	.41**	.48**							
8. Behavioral dis	−.50**	−.49**	−.40**	−.53**	−.44**	−.48**	−.65**						
9. Grades	.21**	.36**	.24**	.23**	.35**	.46**	.22**	−.24**					
10. Female	.03	.13*	.20**	.05	.03	.10	.001	−.003	.12				
11. Black	−.03	−.10	−.04	.01	−.07	−.06	−.05	.01	−.19**	−.10			
12. Asian	.04	−.04	−.04	−.01	.07	.05	.06	−.10	.08	.01	−.09		
13. Hispanic	−.01	−.02	−.04	−.10	−.03	−.09	−.005	.01	−.22**	−.06	−.11	−.07	
14. Multi	.11	.05	−.005	.15*	.06	.01	.07	−.05	.10	.08	−.15*	−.09	−.11
M	2.92	3.69	5.46	3.18	4.45	5.13	3.30	2.44	84.06				
SD	.71	.58	1.17	.72	1.08	1.17	.41	.56	8.49				
Reliability coefficient	.67	.71	.58	.87	.89	.82	.87	.80					

Note. We used pairwise deletion to deal with missing data; *N*s ranged from 190 to 203. Means and standard deviations are based on summed scores that assume that the item responses are interval scales. However, the Likert item responses were treated as categorical variables in further analyses, and expected a posteriori (EAP) scores were used in multiple regression analyses to answer Research Question 3. We did not report means and standard deviations for demographic variables because they are categorical. Bivariate correlations are between EAP scores for all grit, personality, self-regulation, and engagement variables. For grit, these EAP scores were calculated from the best-fitting model for high school, the two correlated-factor model. Because the best-fitting model for the college sample was the bifactor model, we could not compare correlations across age levels. The bivariate correlations among EAP scores are different from the corresponding correlations among latent factors in the multidimensional confirmatory factor analysis models because models used to calculate the EAP scores were unidimensional models that did not take the other correlated constructs into account. The reliability coefficients are for the EAP scores for each variable. Given that there are only two items for conscientiousness, the reliability coefficient of .58 is acceptable. CI = consistency of interests; PE = perseverance of effort; reg = regulation; eng = engagement; dis = disaffection.
* $p < .05$. ** $p < .01$.

scaled surveys (such as the Grit-S and the other measures given in this study) is whether the data should be treated as continuous (i.e., assuming equal intervals between each response option on the scale) or categorical (i.e., assuming unequal intervals between each response option on the scale; Jamieson, 2004). If researchers treat data as continuous when they are actually categorical their analysis may yield less reliable results (Cohen, Manion, & Morrison, 2000; Rhemtulla, Brosseau-Liard, & Savalei, 2012). If data are categorical in nature, researchers should use item factor analysis models, also known as MIRT models for confirmatory item factor analysis, rather than confirmatory factor analysis (CFA) models with continuous item responses (e.g., Wirth & Edwards, 2007).

In a simulation study, Rhemtulla et al. (2012) found that treating categorical variables as continuous variables yielded unbiased or acceptable levels of parameter estimates and standard errors only when the number of categories was at least five and thresholds of item responses were symmetric. When the shape of the response distribution was not symmetric, the equal distance between adjacent categories could not be assumed. The descriptive statistics for participants' responses to the items on the Grit-S (see Figure 1) suggest that participants in both samples did not use all five categories of the response scales for Grit Items 7 and 8. Further, the distributions for nearly all items were not normal. We directly tested the normality assumption for each item response for both samples using Kolmogorov–Smirnov (Smirnov, 1948) and Shapiro-Wilk (Shapiro & Wilk, 1965) tests. The results indicated none of the eight item responses were normally distributed in either population (see Table S1 in the online supplemental materials). Both these pieces of evidence suggest that it is more appropriate to treat the responses to the Grit-S as categorical rather than continuous variables.

Item factor models and estimation. Another important consideration with respect to item response models is what estimation approach to use. Sample size is one issue to consider when making this choice (Forero & Maydeu-Olivares, 2009). Because of our relatively small sample sizes, we used full information maximum likelihood estimation with the expectation–maximization computation algorithm (Bock & Aitkin, 1981). We used the following goodness of model fit indices: (a) an absolute model fit index (M2 statistic root mean square error of approximation [RMSEA]; Joe & Maydeu-Olivares, 2010) and (b) relative model fit indices (observed log-likelihood, Akaike information criterion [AIC], and Bayesian information criterion [BIC]). All of these are available from multidimensional item response theory software flexMIRT (Cai, 2012). See Appendix A in the online supplemental materials for more technical details.

Analyses testing the research questions.

Research Question 1: What is the best-fitting factor structure model of grit in high school and college students?

We used MIRT, also known as confirmatory item factor analysis (Wirth & Edwards, 2007), to answer this research question. To assess the underlying structure of the Grit-S, we tested in each sample the three competing factor models (see Figure 2): a one-factor model that assumes that grit is a single latent construct with

Table 3
Correlations, Means, Standard Deviations, and Reliability Coefficients for all Variables—College Sample

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Grit total														
2. Grit-CI	.40**													
3. Grit-PE	.22**	−.22**												
4. Conscientiousness	.60**	.28**	.35**											
5. Self-control	.67**	.30**	.29**	.71**										
6. Cognitive self-reg	.35**	.15*	.27**	.27**	.35**									
7. Effort reg	.41**	.18**	.38**	.45**	.49**	.57**								
8. Behavioral eng	.40**	.16**	.34**	.38**	.42**	.44**	.43**							
9. Behavioral dis	−.50**	−.35**	−.19**	−.42**	−.51**	−.39**	−.50**	−.65**						
10. Grades	.16**	.04	.21**	.22**	.18**	.26**	.27**	.27**	−.23**					
11. Female	.10	.01	.20**	.18**	.10	.02	.03	.05	.07	.16*				
12. Black	.03	−.06	.01	.06	.09	−.09	.02	−.02	.13*	−.14*	.07			
13. Asian	.06	−.04	−.05	−.05	−.03	.02	.02	−.05	.03	.09	−.04	−.15**		
14. Hispanic	.06	.13*	−.08	−.02	.01	.03	.03	.12*	−.10	−.06	.03	−.09	−.13*	
15. Multi	−.03	−.04	.09	.04	.001	−.07	−.07	−.04	.08	−.13*	.05	−.09	−.13*	−.08
Mean	3.31	2.87	3.76	5.44	3.12	4.51	5.14	3.27	2.37	10.42				
Standard deviation	.63	.82	.64	1.23	.75	.91	1.10	.48	.56	2.08				
Reliability coefficient	.72	.41	.65	.65	.89	.88	.79	.79	.81					

Note. We used pairwise deletion to deal with missing data; Ns ranged from 314 to 336. Means and standard deviations are based on summed scores that assume that the item responses are interval scales. However, the Likert item responses were treated as categorical variables in further analyses and expected a priori (EAP) scores were used in multiple regression analyses to answer Research Question 3. We did not report means and standard deviations for demographic variables because they are categorical. Bivariate correlations are between EAP scores for all variables. For grit, these EAP scores were calculated from the best-fitting model for college, the bifactor model. Because the best-fitting model for the high school sample was the two-correlated factor model, correlations cannot be compared across age levels. Grit-CI and Grit-PE scores from the bifactor model are residual scores relative to general grit factor, and thus Grit-CI and Grit-PE are negatively correlated if the total grit factor is controlled for. The bivariate correlations among EAP scores are different from the corresponding correlations among latent factors in the multidimensional confirmatory factor analysis models because models used to calculate the EAP scores were unidimensional models that did not take the other correlated constructs into account. The reliability coefficients are for the EAP scores for each variable. Given that there are only two items for conscientiousness, the reliability coefficient of .65 is acceptable. The reliability coefficients of Grit-CI and Grit-PE are based on the bifactor model. As these two subscores capture the residual scores after the general grit factor explains variance and only four items load on each factor, the coefficients seem low. But when the scores are used along with the general grit factor in the analysis, the subscores explain more variance above and beyond the general grit factor. Accordingly, lower reliabilities for these subscores are not a critical issue as long as they are used along with general grit scores. CI = consistency of interests; PE = perseverance of effort; reg = regulation; eng = engagement; dis = disaffection.
* $p < .05$. ** $p < .01$.

no subscales, a two correlated-factor model that assumes the proposed subscales of grit are separate, yet correlated, latent constructs, and a bifactor model that assumes grit is a single latent construct with two orthogonal secondary factors, consistency of interests and perseverance of effort, that capture the residual dependency among the items.

To examine whether the high school and college samples were invariant, we followed guidelines by Steenkamp and Baumgartner (1998). First we tested configural invariance by examining whether the factor structure was the same across the two age groups. As will be discussed in more detail subsequently, configural invariance was not established, and thus we did not examine metric or scalar invariance.

> Research Question 2: How empirically distinct or overlapping is grit from conscientiousness, self-control, cognitive self-regulation, effort regulation, behavioral engagement, and behavioral disaffection?

To address this question we ran a series of four MIRT models that included items from the Grit-S and items from one of the other constructs reviewed above (see Figure 3), and assessed comparative model fit. We ran these models with two constructs at a time (e.g., grit and self-control, grit and conscientiousness, etc.; see Christensen & Knezek, 2014, for a similar approach, and Appendix A in the online supplemental materials for further justification). These models build on those used to answer Research Question 1; because we now include constructs in addition to grit in the factor analyses, it was critical to run all four of these types of models, as discussed the following text.

The models range from ones that specify grit and the other construct are quite distinct, to models specifying complete overlap in the constructs. Specifically, Model 1 assumes that consistency of interests (heretofore referred to as Grit-CI), perseverance of effort (heretofore referred to as Grit-PE), and the comparison construct (conscientiousness, self-control, cognitive self-regulation, effort regulation, behavioral engagement, or behavioral disaffection; heretofore referred to as Construct X) are all empirically separate, but correlated, constructs. Model 2 assumes that Grit-CI and Grit-PE make up a single construct called *Grit* (i.e., a bifactor model of grit) and that Grit is correlated with Construct X. Thus, both Models 1 and 2 assume that grit is separate from, yet correlated with, Construct X; in these models, there is little empirical overlap between Grit-S items and items from the measure of Construct X.

Moving to models specifying more overlap, Model 3 assumes that Grit-CI and Construct X make up a single construct (i.e., a bifactor model of Grit-CI and Construct X), and that this construct is separate from but correlated with Grit-PE. Model 4 assumes that Grit-PE and Construct X make up a single con-

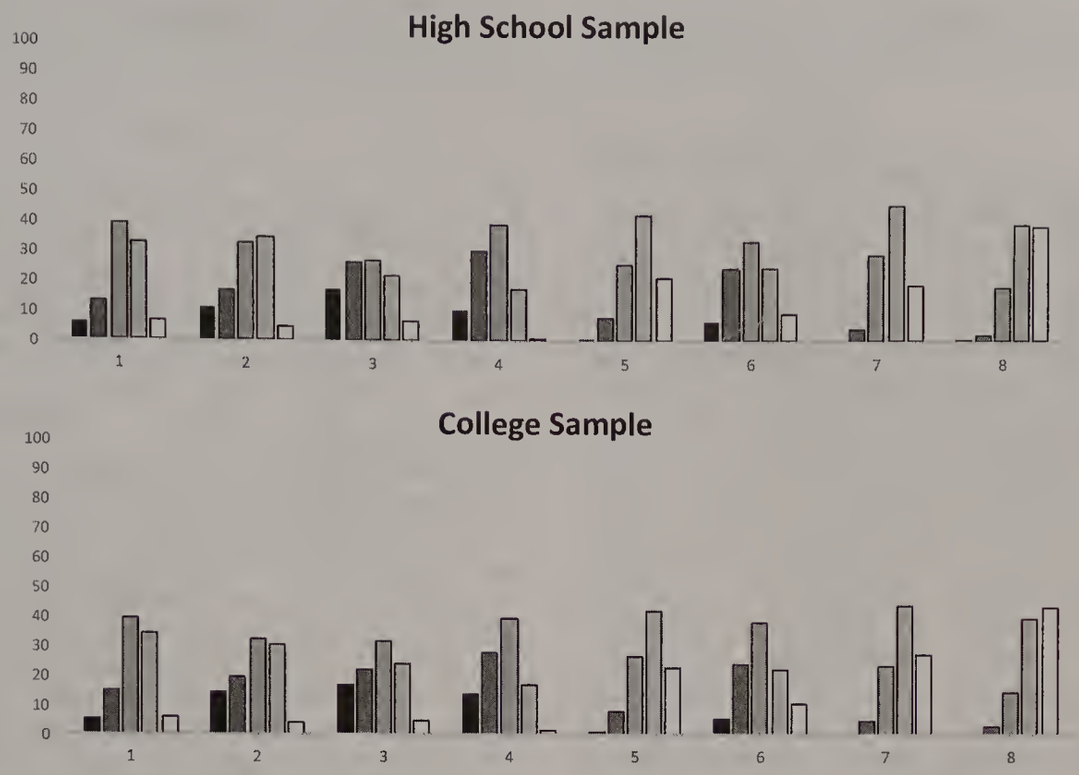


Figure 1. Percentage frequencies of item responses to Grit-S items. None of the responses are normally distributed in population based on the results from two tests of normality (see Table S1 in the online supplemental materials).

struct (i.e., a bifactor model of Grit-PE and Construct X), and that this construct is separate from but correlated with Grit-CI. Thus, Models 3 and 4 decompose grit by assessing whether one of the subscales of grit empirically overlaps with the comparison construct, while the other subscale of grit remains separate. If we find that Models 3 or 4 fit better than Models 1 or 2 for a particular construct (based on the AIC and BIC fit indices), we will have evidence that one of the two subscales of grit appears to overlap with that construct.

We predict this will happen more often in the case of Grit-PE than it will with the Grit-CI—thus, we expect that Model 4 will generally fit best. We base this on the close similarity of items on the Grit-PE subscale (e.g., "I finish whatever I begin", "I am a hard worker") and other scales, including the self-control scale (e.g., "I am lazy"), the effort regulation scale (e.g., "I work hard to do well in this class even if I don't like what we are doing") and the behavioral engagement scale (e.g., "In class, I work as hard as I can"). Although we posit that there will still be overlap with Grit-CI, these items (e.g., "I often set a goal but later choose to pursue a different one") seem less conceptually related to the other constructs.

Research questions 3a, 3b, and 3c: Does students' grit predict their later grades after controlling for gender and ethnicity? Which

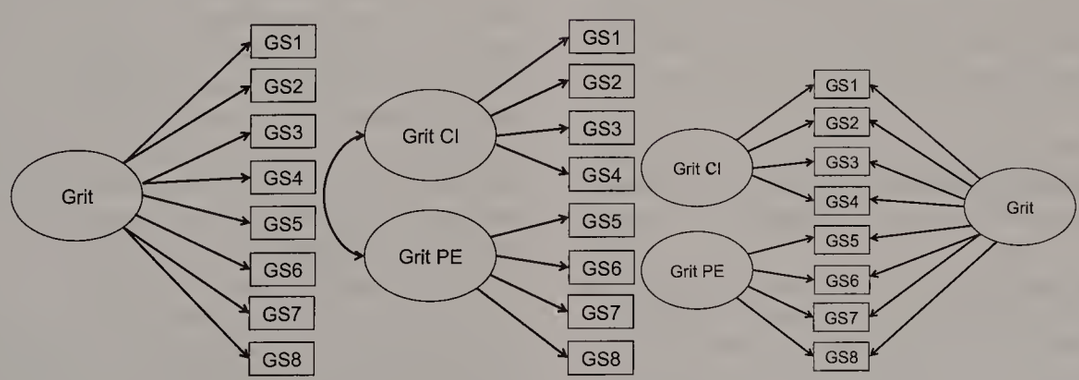


Figure 2. Graphical representations of tested structural models for grit (from left to right: one-factor model, two correlated-factor model, bifactor model). CI = consistency of interests; PE = perseverance of effort.

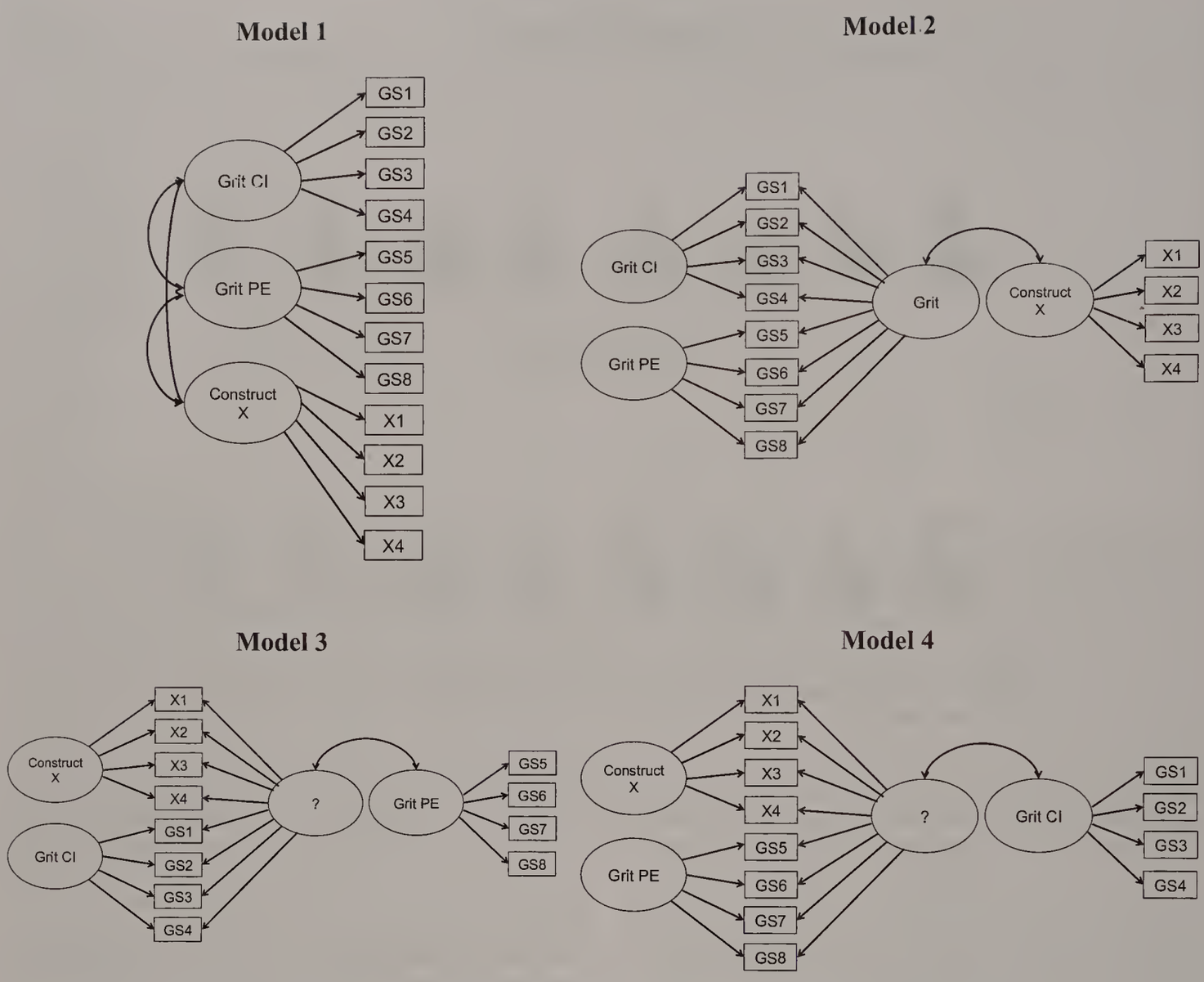


Figure 3. Graphical representations of structural models of grit and conceptually related constructs. CI = consistency of interests; PE = perseverance of effort.

constructs (grit, conscientiousness, self-control, cognitive self-regulation, effort regulation, behavioral engagement, or behavioral disaffection) are the most powerful predictors of grades after controlling for gender and ethnicity? Does students' grit predict their later grades after controlling for gender, ethnicity, and similar constructs?

To examine these questions we chose a two-step approach in which we first obtained latent factor scores for each latent construct from the MIRT analyses and used them in subsequent multiple regression analyses. The factor scores for grit are based on the final structural model we choose to answer Research Question 1 (e.g., if the two correlated-factor model fit best, we used that model to calculate latent factor scores, whereas if the bifactor model fit best, we used that model to calculate latent factor scores). Specifically, we used expected a posteriori (EAP) scores (see Thissen & Wainer, 2001, and Appendix A in the online supplemental materials for more details about using MIRT scale scores).

We conducted a series of multiple regression analyses using the appropriate EAP scores for each construct, including grit. First, to test whether students' grit predicted their later grades after controlling for gender and ethnicity, we regressed students' grades on gender and ethnicity at Step 1 and then added grit EAP scores to the model at Step 2. Second, to test which constructs (grit, conscientiousness, self-control, cognitive self-regulation, effort regulation, behavioral engagement, or behavioral disaffection) were the most powerful predictors of grades after controlling for gender and ethnicity, we ran separate hierarchical regression analyses in which the students' grades are regressed on gender and ethnicity at Step 1, and EAP scores for each construct at Step 2. Standardized partial regression coefficients and changes in R^2 were compared to discuss the effect of each construct on the outcome. Third, in order to test whether students' grit predicted their later grades after controlling for gender, ethnicity, and similar constructs, we conducted a hierarchical regression and regressed students' grades on

gender and ethnicity at Step 1, added the personality, self-regulation, and engagement variables at Step 2, and added grit at Step 3.

Results

Research Question 1: What is the best-fitting factor structure model of grit in high school and college students?

To address Research Question 1, we ran the one-factor model, two correlated-factor model, and bifactor model of grit for each sample. Model fit indices are summarized in Table 4.

High School Sample

For the high school sample, the two correlated-factor model yielded the best fit (see Table 4; also see Appendix B and Table S2 in the online supplemental materials for more details). The likelihood ratio test between the unidimensional model and the two correlated-factor model indicated that the two factor model explained (predicted) the observed eight item response patterns significantly better (–2LL: 4119.64–4061.19 = 58.45, df: 39–38 = 1, $p < .05$). Interestingly, the bifactor models did not properly converge for the high school sample because an item from the persistence of effort subscale ("Setbacks don't discourage me") exhibited a very low and statistically nonsignificant factor loading (0.09, $SE = 0.15$). The discrimination parameter of this item was very low at 0.24 ($SE = 0.18$), indicating that it provided little information about high school students' perseverance of effort. When this item was dropped we found that the model fit indices favored the bifactor model. However, because we wanted to compare our results to those of the previous factor analytic studies of the Grit-S, we decided to retain all of the Grit-S items, including this one. Therefore, we concluded that the two correlated-factor model fit best for the high school sample, and calculated EAP scores (for Grit-CI and Grit-PE separately) based on that factor solution to answer Research Question 3. The correlation estimate between Grit-CI and Grit-PE obtained from the two correlated-factor model was 0.43 ($SE = 0.1$), which is statistically significant (see Table 5b).

Table 4
Model Fit Indices for the Tested Structural Models of the Grit-S

Model	One factor	Two factors	Bifactor
		High school students	
–2 log likelihood	4119.64	**4061.19**	NC
AIC	4197.65	**4146.99**	NC
BIC	4326.67	**4279.32**	NC
No. of parameters	38	39	NC
RMSEA	.06	.06	NC
		College students	
–2 log likelihood	6455.71	6296.96	**6269.44**
AIC	6531.71	6374.96	**6361.47**
BIC	6676.53	**6523.59**	6568.67
No. of parameters	38	39	46
RMSEA	.06	.05	.05

Note. $N = 203$ (high school), $N = 336$ (college). AIC = Akaike information criterion; BIC = Bayesian information criterion; RMSEA = root mean square error of approximation; IRT = item response theory; NC = The model is not properly converged because of nonpositive residual variance for Item 6 and Item 1 for high school and college student samples, respectively. The best-fitting model indices are in boldface.

College Sample

For the college sample, the bifactor model, comprising a primary grit factor and two orthogonal secondary factors (Grit-CI and Grit-PE) that capture the residual dependency among items, fit best (see Table 4; also see Appendix B and Table S2 in the online supplemental materials for more details). The likelihood ratio test between the bifactor model and the other two alternative models showed that the bifactor model was significantly better than the others in explaining the observed item response patterns (–2LL: 6455.71 – 6269.44 = 182.27, df: 46–38 = 8, $p < .05$; –2LL: 6296.96 – 6269.44 = 27.52, df: 46–39 = 7, $p < .05$). Based on these results for the college sample we calculated EAP scores for general grit, Grit-CI, and Grit-PE separately and used these scores in the analysis to address Research Question 3. Tables 2 and 3 present the reliability coefficients for the MIRT scale scores for the high school and college samples. Appendix B in the online supplemental materials provide more technical details about these reliability coefficients and factor loadings for the final grit structural models for both samples.

It should be noted that the EAP scores between high school and college samples are not comparable, because the EAP scores were calculated based on the best-fitting model for each sample, and the high school and college samples had different best-fitting models. For the high school sample, Grit-CI and Grit-PE EAP scores represent the level of Grit-CI and Grit-PE as they are regardless of total Grit level (since the two correlated-factor model fit best). For the college sample, Grit-CI and Grit-PE EAP scores are residual scores after factoring out general Grit scores (since the bifactor model fit best). Therefore, Grit-CI and Grit-PE EAP scores for the college sample need to be interpreted as the levels of Grit-CI and Grit-PE controlling for total Grit level. Tables 2 and 3 include correlations between the EAP scores for all variables for both samples. As the EAP scores between high school and college samples are not comparable, the correlations reported in Tables 2 and 3 are not comparable across the samples. However, comparisons of correlations within each table are appropriate.

Given that the underlying factor structure of grit was different across the two age groups when using the full Grit-S for both groups, the configural invariance condition was not met (Brown, 2006). Thus, stronger measurement invariance tests (e.g., metric invariance test using a multiple group analysis) were not necessary, and we concluded that measurement invariance should not be assumed between the two age groups (Steenkamp & Baumgartner, 1998).

Research Question 2: How empirically distinct or overlapping is grit from conceptually similar constructs?

To address this question, we ran the series of four MIRT models described above for each pair of constructs; these are illustrated in Figure 3. We first discuss overall correlations between grit (or its subscales) and each construct, which were calculated from Models

1 and 2.[4] Then we examined which models (1 through 4) fit best for each pair of constructs (see Tables 5, 6, 7, and 8).

High School Sample

Model 1 posited that Grit-CI, Grit-PE, and Construct X were all separate but correlated latent variables. Thus, from this model we obtained correlations among the three latent variables (i.e., Grit-CI, Grit-PE, and Construct X), which ranged from moderate to high (see Table 6).[5] The correlations between Grit-CI and Grit-PE ranged from 0.42 to 0.43. Importantly, for all of the constructs except for conscientiousness, the correlation between Grit-PE and Construct X was higher than the correlation between Grit-CI and Grit-PE (i.e., the two grit subscales).

Model 2 posited that Grit-CI and Grit-PE made up a larger construct called general grit (i.e., a bifactor model of grit) and that general grit was correlated with Construct X. The correlations between the single general grit factor and each construct indicated moderate to strong overlap in the constructs. The overlap between behavioral disaffection and general grit was particularly high since more than 80% of the variance was shared between behavioral disaffection and general grit.

After fitting Models 1 through 4 for each pair of constructs (i.e., grit and Construct X), we compared which models fit best using AIC and BIC values, because the models were not nested within each other (Burnham & Anderson, 1998). Since the BIC index is less sensitive in evaluating complex models when the sample size is relatively small (such as the samples here; see Yang, 2005), we focus on the AIC values when the AIC and BIC fit values diverge.

Model 3 fit best for effort regulation and behavioral disaffection, which indicates there is one primary factor underlying Grit-CI and these constructs. The primary factor underlying Grit-CI and either effort regulation or behavioral disaffection is also highly correlated with Grit-PE. Similarly, Model 4 fit best for self-control, cognitive self-regulation, and behavioral engagement, which indicates there is one primary factor underlying Grit-PE and each of these constructs. The primary factor underlying Grit-PE and either self-control, cognitive self-regulation, or behavioral engagement is also highly correlated with Grit-CI. Overall, the fact that Models 3 and 4 overall fit better than Models 1 and 2 means that there was a great deal of overlap between the subscales of grit and each comparison construct.

College Sample

Tables 7 and 8 present the results for the college student sample. In Model 1, correlations among the three latent variables (Grit-CI, Grit-PE, and Construct X) ranged from moderate to high. The correlations between Grit-CI and Grit-PE ranged from 0.55 to 0.58, which were slightly higher than for the high school sample. For Model 2, the correlations between the single general grit factor and each construct were moderate to large.

When comparing model fit, as in the high school sample, Models 3 and 4 fit the data better than Models 1 and 2 for all comparison constructs, indicating that there is substantial overlap of grit with the other constructs. Similar to the high school sample, Model 3 fit best for effort regulation and behavioral disaffection (indicating high overlap with Grit-CI) and Model 4 fit best for self-control (indicating high overlap with Grit-PE). However, in the college sample Model 3 fit best for cognitive self-regulation and behavioral engagement (indicating more overlap with Grit-CI than Grit-PE), results that are different from those in the high school sample and contrary to our expectations. Additionally, Model 4 fit best for conscientiousness in the college sample (indicating high overlap with Grit-PE), whereas the models with conscientiousness did not converge for the high school sample.

In summary, across both samples these results suggest that there is a great deal of overlap between grit and the comparison constructs. In fact, the two components of grit appeared to be more strongly related to other constructs in the personality, engagement, and self-regulation literatures than they were to each other. Additionally, the items from the two components of grit factored together with items from other constructs better than they did with each other.

> Research Questions 3a, 3b, and 3c: Does students' grit predict grades after controlling for gender and ethnicity? Which constructs are the most powerful predictors of grades after controlling for gender and ethnicity? Does students' grit predict grades after controlling for gender, ethnicity, and similar constructs?

High School Sample

We ran a series of multiple regression analyses to examine if students' grit (measured in the middle of the semester) predicts their grades (measured at the end of the semester). As noted earlier, EAP scores for grit were calculated based on the best-fitting factor model for Research Question 1. Because the two correlated-factor model fit best for high school students, there were two EAP scores for grit in this sample, Grit-CI and Grit-PE. To answer Research Question 3a, we regressed high school students' grades on gender and ethnicity and found that Black and Hispanic status were negatively associated with student grades; grades did not differ by gender. When Grit-CI and Grit-PE EAP scores were added to the model, an additional 11% of variance in students' grades was explained and the increment in R^2 was statistically significant (see Table 9). However, only Grit-PE was a significant predictor, not Grit-CI.

To answer Research Question 3b, grit and the comparison constructs were entered individually into eight separate hierarchical regression models to see which constructs predicted students' grades at Step 2 when controlling for gender and ethnicity at Step 1 (see Table 10). We then compared the standardized betas and changes in R^2 to make inferences about which constructs were the strongest predictors. Each added construct explained a significant amount of additional variability in students' grades, with effort regulation, Grit-PE, and cognitive self-regulation emerging as the strongest predictors.

---

[4] Note that although Models 1 and 2 did not fit best for any of the comparison constructs, it is still appropriate to look at correlations that were calculated based on these models because all converged models yielded acceptable levels of absolute model fit indices (i.e., values of RMSEA using M2 statistics that are smaller than 0.05).

[5] Note that these correlations were parameters estimated from these models, and therefore different from the EAP score correlations reported in Tables 2 and 3 that came from the models tested in Research Question 1.

Table 5
Model Fit Indices From Confirmatory Factor Analysis for the High School Sample

	Model 1	Model 2	Model 3	Model 4
Self-control				
AIC	11330.33	NC	11171.42	**11169.02**
BIC	11684.32	NC	11575.03	**11572.62**
Conscientiousness				
AIC	NC	NC	NC	5268.74
BIC	NC	NC	NC	5460.91
Cognitive self-regulation				
AIC	11009.70	11001.69	10991.78	**10989.56**
BIC	11380.78	11392.65	11402.62	**11400.40**
Effort regulation				
AIC	6628.94	6632.17	**6616.37**	6617.71
BIC	6860.86	6880.66	**6868.18**	6869.51
Behavioral engagement				
AIC	5873.96	5877.91	5865.05	**5863.05**
BIC	6076.06	6099.90	6090.35	**6088.35**
Behavioral disaffection				
AIC	6192.10	6200.10	**6133.00**	6135.18
BIC	6397.52	6425.39	**6361.61**	6363.79

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; NC = estimation is not converged. The best fitting model indices are bolded.

Finally, to answer Question 3c, we conducted a hierarchical regression and entered demographic variables at Step 1, personality, self-regulation, and engagement constructs at Step 2, and Grit-CI and Grit-PE at Step 3. The change in F was significant at Step 1 ($p = .001$), and Step 2 ($p < .001$), but not at Step 3 ($p = .24$). About 30% of the variance in grades was explained by all variables at Step 3 (see Table 11). The strongest predictor was effort regulation; this was the only statistically significant predictor at Step 3. The next strongest predictor based on the standardized betas was Grit-PE, although it was not statistically significant. Thus, grit did not significantly predict grades above and beyond this set of demographic, personality, self-regulation, and engagement variables.

College Sample

In this sample, participants' grit EAP scores from the bifactor model (see Figure 2) were utilized, as the bifactor model fit the data best. Thus, there were three EAP scores: general grit, Grit-CI, and Grit-PE.[6] As with the high school students, to answer Research Question 3a, we regressed college students' grades on gender and ethnicity and found that females obtained higher grades and Black and multiracial students lower grades (see Table 9). General grit was added next and it was a significant predictor of students' grades. When the other two subscale grit EAP scores (Grit-CI and Grit-PE) from the bifactor model were entered into the regression model, only Grit-PE significantly predicted students' grades; these results are similar to the high school results.

To answer Research Question 3b, we examined the predictive relations of each individual construct to students' grades by entering the constructs into nine separate regression models (controlling for gender and ethnicity), and comparing the standardized betas. Results are summarized in Table 10. Grit-CI was not a statistically significant predictor of students' grades, but all the other constructs were. Similar to the results for the high school students, effort regulation was one of the strongest predictors, along with behavioral engagement and cognitive self-regulation.

To answer Research Question 3c, we conducted a hierarchical regression and entered demographic variables at Step 1, personality, self-regulation, and engagement constructs at Step 2, and general grit, Grit-CI and Grit-PE at Step 3. The change in F was significant at Step 1 ($p < .001$), and Step 2 ($p < .001$), but not at Step 3 ($p = .80$). None of individual psychological (i.e., nondemographic) predictors were statistically significant at Step 3 (see Table 12). The overall model explained about 18% of the variance in students' grades. The strongest predictor above and beyond the demographic variables based on the standardized betas was behavioral engagement, followed by conscientiousness, cognitive self-regulation, and effort regulation, although none of these were statistically significant.

In summary, in both samples neither Grit-CI nor Grit-PE were significant predictors above and beyond the demographic, personality, self-regulation, and engagement variables. Based on the size of the standardized betas, Grit-PE was a stronger predictor of students' grades than Grit-CI for both high school and college students. However, in both samples effort regulation and behavioral engagement predicted students' grades more strongly than did the grit variables.

[6] EAP scores for general grit, Grit-CI, and Grit-PE from a bifactor model should be carefully interpreted since the factors are orthogonal (i.e., uncorrelated) in the model. The shared variance among all items is captured by general grit. After partialing out the general factor, uniquely shared variance or residual dependency among only Grit-CI items is captured by the specific factor called Grit-CI, and uniquely shared variance among only the Grit-PE items is captured by another specific factor called Grit-PE. Accordingly, the correlation coefficients among EAP scores are not high enough to cause a multicollinearity issue.

Table 6
Correlations Among Factors From Confirmatory Factor Analysis for the High School Sample

	Model 1			Model 2	Model 3	Model 4
	Grit-CI-X	Grit-PE-X	Grit-CI-PE	Grit-X	FacX-Grit-PE	FacY-Grit-CI
Self-control	.75 (.06)	.65 (.07)	.42 (.10)	NC	.76 (.06)	**.75 (.06)**
Conscientiousness	NC	NC	NC	NC	NC	.46 (.09)
Cognitive self-regulation	.35 (.09)	.55 (.07)	.43 (.10)	.62 (.08)	.61 (.08)	**.63 (.09)**
Effort regulation	.42 (.09)	.75 (.06)	.42 (.10)	.76 (.06)	**.77 (.06)**	.42 (.09)
Behavioral engagement	.36 (.10)	.87 (.06)	.43 (.09)	.87 (.06)	.90 (.07)	**.40 (.09)**
Behavioral disaffection	−.63 (.07)	−.65 (.08)	.42 (.10)	−.97 (.09)	**−.73 (.07)**	−.57 (.08)

Note. N = 203 (high school), 336 (college). CI = consistency of interests; PE = perseverance of effort; X = construct X; FacX = factor underlying construct X and Grit-CI; FacY = factor underlying construct X and Grit-PE; NC = estimation is not converged. Correlation estimates were obtained from CFA results. Standard errors are in parentheses. Correlations from the best-fitting models are bolded.

Discussion

Grit has become a much discussed construct in the research literature and the popular press (Tough, 2012; Sehgal, 2015), and government agencies have incorporated it into recommendations for how practitioners might work with children to help them develop the "non-cognitive" skills they need to succeed in school (U.S. Department of Education, 2013). All of this has occurred even though work examining the construct validity of grit, its relations to other seemingly similar constructs, and its relations to achievement outcomes is relatively sparse and contains conflicting findings. In the present study we examined each of these issues in samples of high school and college students; we begin this section discussing what we found regarding grit's factor structure.

The Factor Structure of Grit in High School and College Students

To investigate further whether grit emerged as a single construct with two subscales as Duckworth and colleagues (2007) hypothesized and found, we assessed grit's factor structure (utilizing CFA with categorical variables, something not often done in this literature) in our high school and college samples. We tested three different models: a one factor model, a two correlated-factor model that posited consistency of interests (Grit-CI) and perseverance of effort (Grit-PE) as two separate latent constructs, and a bifactor model that posited grit as a single latent construct with two subscales (Grit-CI and Grit-PE). In our college sample the bifactor model fit best, and in our high school sample the two-correlated model fit best. This is an important extension of Duckworth et al. (2007), Duckworth and Quinn (2009), and others' work on grit's factor structure, which has not examined a bifactor model. As discussed above, the higher order factor model tested in previous work is equivalent to a two correlated-factor model when there are only two factors or subscales. The bifactor model is more appropriate morel to test if grit is a single construct with two scales as originally conceptualized by Duckworth and colleagues (2007). Results for the college students provide support for this conceptualization of grit.

The best-fitting model differed to some degree across the high school and college samples; however, this was due primarily to the perseverance of effort item "Setbacks don't discourage me" having

Table 7
Model Fit Indices From Confirmatory Factor Analysis for the College Sample

	Model 1	Model 2	Model 3	Model 4
Self-control				
AIC	17797.83	17784.88	17612.30	**17602.76**
BIC	18202.13	18212.06	18073.81	**18064.27**
Conscientiousness				
AIC	8242.17	NC	8227.15	**8215.93**
BIC	8448.14	NC	8448.37	**8437.15**
Cognitive self-regulation				
AIC	19545.86	19536.33	**19491.83**	20355.83
BIC	20022.26	20035.59	**20021.57**	20962.75
Effort regulation				
AIC	10468.90	10470.67	**10461.97**	10464.55
BIC	10731.87	10756.51	**10747.81**	10750.39
Behavioral engagement				
AIC	9147.07	9145.89	**9073.88**	9076.45
BIC	9379.91	9401.64	**9333.44**	9336.01
Behavioral disaffection				
AIC	9704.62	9699.25	**9630.15**	9632.26
BIC	9937.28	9954.80	**9889.51**	9891.62

Note. AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion; NC = estimation is not converged. The best fitting model indices are bolded.

Table 8
Correlations Among Factors From Confirmatory Factor Analysis for the College Sample

	Model 1			Model 2	Model 3	Model 4
	Grit-CI-X	Grit-PE-X	Grit-CI-PE	Grit-X	FacX-Grit-PE	FacY-Grit-CI
Self-control	.76 (.04)	.68 (.04)	.58 (.05)	.92 (.04)	.74 (.04)	**.81 (.03)**
Conscientiousness	.83 (.07)	.85 (.07)	.55 (.05)	NC	.77 (.05)	**.77 (.04)**
Cognitive self-regulation	.37 (.06)	.47 (.05)	.58 (.05)	.54 (.06)	**.62 (.06)**	.91 (.05)
Effort regulation	.46 (.06)	.64 (.05)	.57 (.05)	.68 (.06)	**.97 (.07)**	.60 (.06)
Behavioral engagement	.47 (.06)	.58 (.05)	.57 (.05)	.65 (.06)	**.86 (.04)**	.66 (.05)
Behavioral disaffection	−.66 (.05)	−.52 (.06)	.57 (.05)	−.69 (.06)	**−.73 (.07)**	−.69 (.05)

Note. N = 203 (high school), 336 (college). CI = consistency of interests; PE = perseverance of effort; X = construct X; FacX = factor underlying construct X and Grit-CI; FacY = factor underlying construct X and Grit-PE; NC = estimation is not converged. Correlation estimates were obtained from CFA results. Standard errors are in parentheses. Correlations from the best-fitting models are bolded.

an extremely low factor loading in the high school sample, which led to the bifactor model not properly converging. When this item was not included in the analyses, the bifactor model fit best for high school students. Future studies should examine this item's loading carefully. Pending further examinations of this item in other high school samples, the evidence in this study supports the conceptualization of grit as a single construct with two scales, as Duckworth and colleagues (Duckworth et al., 2007; Duckworth & Quinn, 2009) proposed. However, the reliabilities of the two subscales for both samples were not particularly high (see Tables 2 and 3), as also has been found in some other research (e.g., Duckworth & Quinn, 2009). In the future, researchers should continue to address grit's factor structure, its reliability, and the correlations of the proposed subscales for different age groups, and make modifications to the measure if needed.

A critical methodological finding from our analyses of the items on the Grit-S is that it is not appropriate to treat students' responses to the grit measure as continuous variables. We directly tested the normality assumption and found that it did not hold for any of the Grit-S items. Additionally, for several of the items participants used fewer than five categories. Thus, neither of the assumptions of Rhemtulla et al. (2012) were met, indicating that it was not appropriate to treat the data as continuous. To date, no researcher examining the factor structure of grit has tested whether it is viable to assume that the indicators were continuously distributed; indeed, few if any researchers studying grit, self-regulation, or engagement have tested this assumption. As Rhemtulla and colleagues (2012) explained, incorrectly assuming variables are continuous can yield underestimated factor loadings and parameter standard errors, resulting in mistaken conclusions about the nature of the factor structure for a given measure. In the present study, the bifactor model would not been chosen for the college sample (see Table 4) had continuous indicators been assumed. Further, the subsequent analyses looking at how grit predicted student achievement would have likely underestimated the relations among the constructs and outcomes. Researchers need to check this assumption carefully in subsequent work as the strength of relations among constructs and outcomes has many implications for both our understanding of the construct and educational implications concerning grit.

Table 9
Grit's Prediction of Grades After Controlling for Gender and Ethnicity

	High school						College								
	Step 1			Step 2			Step 1			Step 2			Step 3		
	B	*β̂	*SE*	B	*β̂	*SE*	B	*β̂	*SE*	B	*β̂	*SE*	B	*β̂	*SE*
Female	1.38	.08	1.21	.68	.04	1.15	.84**	.18	.26	.76**	.16	.26	.61*	.13	.26
Black	−4.94*	−.20	1.81	−4.17*	−.17	1.72	−1.17**	−.16	.4	−1.20**	−.17	.4	−1.16**	−.16	.4
Asian	2.16	.06	2.83	2.87	.07	2.68	.23	.04	.31	.17	.03	.3	.20	.04	.3
Hispanic	−7.35**	−.23	2.32	−7.06**	−.22	2.19	−.68	−.09	.45	−.76	−.1	.44	−.64	−.08	.44
Multi	1.12	.04	1.84	.99	.04	1.75	−1.23*	−.15	.47	−1.22*	−.15	.46	−1.32**	−.16	.46
Grit										.38*	.15	.14	.27	.11	.16
Grit-CI				−.02	−.002	.88							.10	.03	.2
Grit-PE				3.44**	.34	.85							.43*	.17	.15
R^2	.11			.22			.08			.10			.12		
Adjusted R^2	.08			.19			.06			.08			.10		
df	184			184			313			313			313		
F ratio	4.28 (p = .001)			7.02 (p < .001)			5.14 (p < .001)			5.66 (p < .001)			5.35 (p < .001)		

Note. N = 185 (high school), N = 314 (college). B = unstandardized regression coefficients; *β̂ = standardized regression coefficients; CI = consistency of interests; PE = perseverance of effort. Because the two correlated-factor model fit best for the high school sample, there was no general grit expected a prioriscore.
* p < .05. ** p < .01.

Table 10
Predictions of Students' Grades by Each Construct Individually (Controlling for Gender and Ethnicity)

	High school							College						
	B	*β̂	*SE*	ΔR^2	Adj. R^2	*df*	*F* change	B	*β̂	*SE*	ΔR^2	Adj. R^2	*df*	*F* change
Baseline model				.11	.08	179	4.28 ($p = .001$)				.08	.06	308	5.14 ($p < .001$)
Grit								.38*	.15	.14	.02	.08	307	7.69 ($p = .01$)
Grit-CI	2.06*	.20	.74	.04	.12	178	7.76 ($p = .01$)	.12	.04	.18	.00	.06	307	.44 ($p = .51$)
Grit-PE	3.43**	.34	.68	.11	.19	178	25.15 ($p < .001$)	.47**	.18	.14	.03	.09	307	10.79 ($p = .001$)
Conscientiousness	2.55**	.22	.80	.05	.13	178	10.09 ($p = .002$)	.55**	.21	.15	.04	.10	307	14.20 ($p < .001$)
Self-control	1.98**	.21	.67	.04	.12	178	8.76 ($p = .004$)	.39**	.18	.12	.03	.09	307	10.43 ($p = .001$)
Cognitive self-reg	3.09**	.33	.63	.11	.19	178	23.98 ($p < .001$)	.53**	.24	.12	.06	.12	307	19.82 ($p < .001$)
Effort reg	4.17**	.43	.62	.18	.26	178	44.84 ($p < .001$)	.61**	.25	.13	.06	.12	307	22.80 ($p < .001$)
Behavioral eng	2.09**	.20	.73	.04	.12	178	8.34 ($p = .004$)	.60**	.26	.12	.07	.13	307	24.08 ($p < .001$)
Behavioral dis	−2.31**	−.24	.66	.06	.14	178	12.36 ($p = .001$)	−.54**	−.23	.13	.05	.11	307	18.12 ($p < .001$)

Note. $N = 185$ (high school), $N = 314$ (college). Each construct was alternately tested while gender and ethnicity are consistently remained in the regression model. The "baseline model" refers to a model with only gender and ethnicity entered; the change in R^2 is after each variable is entered above that baseline model. Because the two correlated-factor model fit best for the high school sample, there was no general grit EAP score. B = unstandardized regression coefficients; *β̂ = standardized regression coefficients; CI = consistency of interests; PE = perseverance of effort; reg = regulation; eng = engagement; dis = disaffection; adj = adjusted.
* $p < .05$. ** $p < .01$.

Grit's Relation to Conceptually and Operationally Similar Constructs

As discussed in the introduction, grit overlaps conceptually and operationally with other constructs based in the personality, self-regulation, and engagement literatures. Yet few studies have compared systematically the empirical overlap of these constructs, with the exception of some studies showing that grit relates strongly to conscientiousness (e.g., Duckworth et al., 2007; Duckworth & Quinn, 2009; Eskreis-Winkler et al., 2014). Thus, a major contribution of the present study is our examination of the overlap and distinctiveness of grit with these constructs.

To address this issue we ran four types of CFA models, two of which assumed that grit was distinct from the other constructs (Models 1 and 2), and two that proposed substantial overlap between them (Models 3 and 4). The first two models did not fit the data as well compared to the other two, and so Models 3 and 4 are the preferred models in this set of analyses. Further, even in Models 1 and 2 the latent correlations of grit and the other constructs were substantial, indicating much overlap between them.

In Models 3 and 4 that tested directly the overlap of grit with individual constructs, high school participants' Grit-CI overlapped most with effort regulation and behavioral disaffection. Effort regulation, which is defined as maintaining effort even when things become difficult or boring, is highly similar to Grit-CI, which focuses on maintaining interests in the face of setbacks or

Table 11
Hierarchical Regression Analysis of Semester Grades With all Constructs Entered as Predictors in the Same Analysis in the High School Sample

	B	*β̂	*SE*	VIF	B	*β̂	*SE*	VIF	B	*β̂	*SE*	VIF
Female	1.49	.09	.98	1.02	.86	.05	1.14	1.12	.93	.06	1.14	1.12
Black	−4.53*	−.18	1.82	1.05	−4.12*	−.16	1.68	1.09	−3.99*	−.16	1.68	1.10
Asian	2.17	.06	2.80	1.02	1.57	.04	2.58	1.05	1.94	.05	2.58	1.06
Hispanic	−7.33**	−.23	2.29	1.04	−6.13**	−.19	2.10	1.06	−6.24**	−.20	2.10	1.06
Multi	1.86	.07	1.96	1.05	2.47	.09	1.82	1.10	2.40	.09	1.82	1.10
Conscientiousness					.54	.05	.92	1.12	−.18	−.02	1.01	2.00
Self-control					−.41	−.04	.80	1.64	−.45	−.05	.85	2.08
Cognitive self-regulation					−.05	−.01	.96	2.57	−.15	−.02	.96	2.59
Effort regulation					4.09**	.43	1.07	3.08	3.72**	.39	1.09	3.21
Behavioral engagement					−.53	−.05	.92	2.06	−.99	−.10	1.00	2.42
Behavioral disaffection					−.55	−.06	.88	2.14	−.06	−.06	.92	2.36
Grit-CI									−.27	−.03	.94	2.02
Grit-PE									1.82	.18	1.11	2.95
ΔR^2		.11				.18				.01		
Adjusted R^2		.08				.24				.25		
df		186				186				186		
F ratio		4.31 ($p = .001$)				6.46 ($p < .001$)				5.71 ($p < .001$)		

Note. $N = 187$. B = unstandardized regression coefficients; *β̂ = standardized regression coefficients; VIF = variance inflation factor; Multi = multi-racial; CI = consistency of interests; PE = perseverance of effort.
* $p < .05$. ** $p < .01$.

Table 12
Hierarchical Regression Analysis of Semester Grades With All Constructs Entered as Predictors in the Same Analysis in the College Sample

	B	*β̂	*SE*	VIF	B	*β̂	*SE*	VIF	B	*β̂	*SE*	VIF
Female	.84**	.18	.26	1.01	.72**	.15	.26	1.08	.69**	.15	.26	1.12
Black	−1.17**	−.16	.40	1.06	−.99*	−.14	.40	1.13	−1.00*	−.14	.40	1.14
Asian	.23	.04	.31	1.07	.23	.04	.29	1.08	.23	.04	.30	1.09
Hispanic	−.68	.13	.45	1.04	−.85*	−.11	.43	1.06	−.78	−.10	.44	1.09
Multi	−1.23**	−.15	.47	1.04	−1.04*	−.13	.45	1.07	−1.08*	−.13	.45	1.08
Conscientiousness					.27	.10	.20	2.08	.27	.10	.21	2.24
Self-control					−.12	−.05	.18	2.29	−.09	−.04	.19	2.61
Cognitive self-regulation					.22	.10	.15	1.63	.22	.10	.15	1.65
Effort regulation					.26	.11	.17	1.88	.22	.09	.18	2.00
Behavioral engagement					.32	.14	.17	1.92	.28	.12	.17	2.02
Behavioral disaffection					−.08	−.04	.18	2.23	−.13	−.02	.19	2.14
Grit									−.06	−.02	.19	2.14
Grit-CI									−.08	−.03	.20	1.50
Grit-PE									.11	.04	.17	1.58
ΔR^2		.08				.10				.003		
Adjusted R^2		.06				.15				.15		
df		313				313				313		
F ratio		5.14 ($p < .001$)				6.07 ($p < .001$)				4.81 ($p < .001$)		

Note. $N = 314$. B = unstandardized regression coefficients; *β̂ = standardized regression coefficients; VIF = variance inflation factor; Multi = multi-racial; CI = consistency of interests; PE = perseverance of effort.
$^{*}p < .05$. $^{**}p < .01$.

distractions. Similarly, behavioral disaffection, which includes being distracted and trying to get out of work, should be negatively related to maintaining interests over time, since individuals who maintain interest in a task typically do not try to avoid it.

High school participants' Grit-PE overlapped most with self-control, cognitive self-regulation, and behavioral engagement. Many of the self-control items are very similar to the Grit-PE, as are many of the cognitive self-regulation items. Items on behavioral engagement scale and the Grit-PE scale are also extremely similar and include items that are almost identical (e.g., "I am a hard worker" on the Grit-PE scale and "I try hard to do well in school" on the behavioral engagement scale). Hence this overlap is not surprising. It is important to note that for high school students, three out of the four models that involved conscientiousness did not converge. This, combined with the low reliability of the measure, suggests that in the future researchers should use a more reliable measure of conscientiousness.

Models 3 and 4 in the college student sample showed that Grit-PE overlapped with self-control and conscientiousness, as would be expected given the conceptual similarities between Grit-PE and these personality constructs. However, cognitive self-regulation, effort regulation, behavioral engagement, and behavioral disaffection overlapped more with Grit-CI than Grit-PE. These results were somewhat surprising given that both conceptually and operationally Grit-PE seems more similar to these other constructs. Perhaps during college when students choose majors and take a series of courses in their major to get a degree, regulatory processes go hand-in-hand with maintaining focus and interests over time. In support of this view, Pintrich and Zusho (2007) discussed the importance of persistence for college students, and maintaining their self-regulatory strategies and interests when encountering increasingly difficult courses and assignments in particular courses (see also Zusho & Edwards, 2011).

As a whole, the results pertaining to Research Question 2 show that there is substantial overlap of grit and its two subscales with various other frequently studied constructs in the personality, self-regulation, and engagement literatures. Thus, grit is not clearly distinguished from these constructs operationally, even if it is conceptually. These results suggest a jangle fallacy (Block, 1995; Whiteside & Lynam, 2001) may be operating; that is, there are different names being given to quite similar constructs. We believe this has occurred because researchers have not systematically examined the relations among these constructs to date. One reason is that the work on these constructs has occurred in different fields (the personality and educational psychology fields), and researchers in these fields publish in different journals and often are not in contact. We believe that this study takes the first step in rectifying this problem.

Although empirically we found much overlap in the various constructs, conceptual distinctions among them still are potentially meaningful. Duckworth and colleagues (2007; Duckworth & Quinn, 2009) distinguished grit from other constructs in the personality literature by focusing on long-term outcomes and the necessity of perseverance of effort and consistency of interests to obtain them. Similarly, researchers in the engagement and self-regulation literatures have distinguished many of these constructs theoretically from one another (see Christenson et al., 2012; Zimmerman & Schunk, 2011 for discussion). One problem is that these conceptual distinctions do not always appear in measures of the constructs. For example, as discussed briefly earlier, the Grit-S measure contains few if any items tapping long-term goals, and what is meant by long-term is not defined. Specifically, these items use the ill-defined term *later* to refer to long-term goals, and one item includes the phrase "projects that take more than a few months to complete." Neither of these are necessarily long-term, especially for college students and adults. Thus, the current scales measuring grit do not fully capture the important theoretical distinctions between being gritty and simply working hard, having self-control, self-regulating, or being engaged in class. It is very important that researchers make sure that measures adequately

capture the theoretical aspects of their construct, and so the Grit-S should be revised to be clearer in the way it assesses long-term goals.

We should note that the jangle fallacy is not only a problem with grit. If we had focused on the factor structure of, for example, the engagement and disaffection measure and looked at its overlap with the other constructs included in this study we likely would be reaching the same conclusions. In fact, we believe our results reflect at least two larger problems in the areas of psychology examining constructs such as the ones studied here: researchers often (a) fail to define a given construct clearly and consistently and (b) develop measures of a given construct that fit its definition and do not include items reflecting other constructs (e.g., see Bong & Skaalvik, 2003; Marsh, Craven, Hinkley, & Debus, 2003; Murphy & Alexander, 2000; Pajares, 1996; Reschly & Christenson, 2012). This issue may be especially problematic when constructs become the focus of intervention work; we return to this point in the following text.

Grit's Prediction of Students' Grades

Results of the present study build on previous work on grit, self-regulation, and engagement showing that each relates to students' grades (Duckworth et al., 2007; Pintrich & De Groot, 1990; Skinner, Wellborn, & Connell, 1990; Zimmerman, 2011). When the two Grit subscales are the only psychological variables entered in the regression equations, they predict 11% of the variance in high school students' grades. In the college student sample general grit and the two subscales predicted significant additional variance above and beyond demographic variables in students' grades, but Grit-PE was the only significant predictor. In both samples when the other psychological variables were included, other variables such as effort regulation and behavioral engagement were stronger predictors of grades than grit.

Thus, for both high school and college students, perseverance of effort predicted grades more strongly than did consistency of interests; this finding is consistent with prior research (e.g., see Chang, 2014). However, in both samples effort regulation consistently predicted students' grades more strongly than grit perseverance of effort did. One explanation for this finding is the "match" of the level of measurement specificity. The perseverance of items are at the domain general level (e.g., I am a hard worker), whereas the effort regulation items are at the specific class level (e.g., I work hard to do well in this class even if I do not like what we are doing). Bandura (1997) and others have argued that self-perception variables relate more strongly to outcomes when they are measured at the same level; given that class grades were our outcome measure, effort regulation provides this match. Also, the effort regulation items describe both effort itself and continuing one's effort even when tasks are uninteresting; being able to do this may be particularly important for outcomes such as grades.

Importantly, our inclusion of this large set of variables gave us a clearer picture of their relative predictive power. Ivcevic and Brackett (2014) reported that grit did not predict high school students' grades when personality variables were controlled; our results support and build on this finding. Perhaps in our study grit was not as predictive as in previous studies because we did not measure the kinds of long-term achievement outcomes that Duckworth and colleagues (2007) posited are predicted by grit. Another possibility is that the results reflect the jangle fallacy alluded to earlier. Grit, effort regulation, cognitive self-regulation, and engagement overlap greatly conceptually and empirically, and so it is not surprising that each explain about the same amount of variance in an important achievement outcome. Future research should address whether grit predicts more variance in individuals' achievement of longer-term outcomes, such as college graduation or overall GPA, than do either self-regulation or engagement for current activities.

The results showing that the variables explained less variance in grades for college than high school students is surprising given that courses in college are generally more demanding and difficult than high school courses, presumably requiring students to have more grit and self-regulation to do well in them. One possibility is that motivational variables, such as the values students hold for different subjects (Wigfield, Tonks, & Klauda, in press), are especially important predictors of students' grades in college. College students have more choice about what courses they take, and these choices (and their performance), thus, may be driven more by how much they value what they are taking (Wang & Eccles, 2012). By contrast, high school students are required to take certain courses; thus they may need to be gritty and otherwise self-regulated to succeed in them, particularly if they do not value them as much.

What do these findings mean for the viability of interventions to improve grit and the current calls in the popular press (Tough, 2012), at federal agencies like IES (U.S. Department of Education, 2013), and in different school districts that grit interventions be developed? The rationale for these calls is that some and possibly many students need to learn to be gritty in order to persist through challenges and difficulties that they face to attain their longer term goals, and once they do so, they will do better in school. To date, we know of no published research that has been done on the effectiveness of such interventions with respect to either short or long-term academic outcomes. However, given the results here concerning relations of grit to grades, we suggest some caution be taken in advocating grit interventions as a way of enhancing school performance. Given that effort regulation, which is more domain-specific, related more strongly to grades than Grit-PE, which is more domain-general, researches should consider the "match" of specificity of construct that is the focus of the intervention and the outcomes they are trying to improve when designing interventions.

Further, we know of no work comparing the effectiveness of interventions focused on grit in comparison to the effectiveness of interventions designed to improve students' self-regulation at different levels of schooling. Some self-regulation interventions have had documented success in improving student academic outcomes (e.g., Zimmerman & Cleary, 2009; Zusho & Edwards, 2011). These programs may be more or less effective depending upon the time frame of the achievement activity to be completed. We think comparing the relative effectiveness of grit interventions to other kinds of previously validated self-regulation interventions is an important next step for research in this area.

Limitations and Other Suggestions for Future Research

The present study examined personality variables including grit at a general level, the engagement variables at the school level, and the self-regulation variables at the level of specific classes. It is

possible that results would be different in different subject areas and for constructs measured at different levels, but that has not been addressed in previous research. In addition, we measured the constructs at one point in time; given the assumption that grit is relatively stable, longitudinal studies should be done to examine how stable it actually is over time and whether its mean level changes over time. Much research suggests that students' motivation for achievement declines across the school years (Wigfield et al., 2015); is the same true for grit? Or, does grit increase as classes get harder? Additionally, longitudinal studies would allow for the interesting examination of possible reciprocal effects between grit, the other variables, and achievement, something that has been found consistently in the literature on relations of self-concept and achievement (Marsh, Martin, Yeung, & Craven, in press). We examined grades as our achievement outcome variable; researchers should examine the relations of the constructs we measured to other kinds of outcomes, particularly ones requiring long-term effort and maintenance of interests.

In conclusion, this study provides important new information about the factor structure of grit and its relations to conceptually similar variables and students' grades. The perseverance of effort component of grit predicts grades more strongly than consistency of interests. Other constructs in the self-regulation, engagement, and personality literatures highly overlap with grit and predict students' grades more strongly than does grit. Based on these results we suggest that researchers look carefully at the measures used to assess constructs like grit and self-regulation, to be sure that the measures reflect accurately the theoretical definitions of the constructs. Perhaps some items on the Grit-S could be modified to reflect more accurately the focus of the construct on more long-term goals. Additionally, policymakers and intervention researchers should choose the variables on which they will intervene (whether grit or others) in terms of their match to the outcomes they hope to improve.

References

Babakus, E., Ferguson, C. E., & Jöreskog, K. G. (1977). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research, 37,* 72–141.

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science, 21,* 372–374. http://dx.doi.org/10.1177/0956797609360756

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York, NY: Freeman.

Block, J. (1995). Going beyond the five factors given: Rejoinder to Costa and McCrae (1995). and Goldberg and Saucier (1995). *Psychological Bulletin, 117,* 226–229. http://dx.doi.org/10.1037/0033-2909.117.2.226

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459. http://dx.doi.org/10.1007/BF02293801

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research, 17,* 303–316. http://dx.doi.org/10.1177/0049124189017003004

Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review, 15,* 1–40. http://dx.doi.org/10.1023/A:1021302408382

Brown, T. A. (2006). *Confirmatory factor analysis for applied research.* New York, NY: Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sage Focus Editions, 154,* 136.

Burnham, K. P., & Anderson, D. R. (1998). *Model selection and inference: A practical information-theoretic approach.* New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4757-2917-7

Cai, L. (2012). *flexMIRT: Flexible multilevel item factor analysis and test scoring* [Computer software]. Seattle, WA: Vector Psychometric Group.

Carney, D. R., Jost, J. T., Gosling, S. D., & Potter, J. (2008). The secret lives of liberals and conservatives: Personality profiles, interaction styles, and the things they leave behind. *Political Psychology, 29,* 807–840. http://dx.doi.org/10.1111/j.1467-9221.2008.00668.x

Chang, W. (2014). *Grit and academic performance: Is being grittier better?* (Doctoral dissertation). Retrieved from http://scholarlyrepository.miami.edu/cgi/viewcontent.cgi?article=2319&context=oa_dissertations

Christensen, R., & Knezek, G. (2014). Comparative measures of grit, tenacity, and perseverance. *International Journal of Learning, Teaching, and Educational Research, 8,* 16–30.

Christenson, S., Reschly, A. L., & Wylie, C. (2012). *Handbook of research on student engagement.* New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-2018-7

Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education* (5th ed.). London, UK: Routledge. http://dx.doi.org/10.4324/9780203224342

Correa, T., Hinsley, A. W., & De Zuniga, H. G. (2010). Who interacts on the Web?: The intersection of users' personality and social media use. *Computers in Human Behavior, 26,* 247–253. http://dx.doi.org/10.1016/j.chb.2009.09.003

Cross, T. M. (2014). The gritty: Grit and non-traditional doctoral student success. *The Journal of Educators Online, 11.* Retrieved from http://files.eric.ed.gov/fulltext/EJ1033306.pdf

de Ridder, D. T., Lensvelt-Mulders, G., Finkenauer, C., Stok, F. M., & Baumeister, R. F. (2012). Taking stock of self-control: A meta-analysis of how trait self-control relates to a wide range of behaviors. *Personality and Social Psychology Review, 16,* 76–99. http://dx.doi.org/10.1177/1088868311418749

Duckworth, A., & Gross, J. J. (2014). Self-control and grit related but separable determinants of success. *Current Directions in Psychological Science, 23,* 319–325. http://dx.doi.org/10.1177/0963721414541462

Duckworth, A. L., Kirby, T. A., Tsukayama, E., Berstein, H., & Ericsson, K. A. (2011). Deliberate practice spells success: Why grittier competitors triumph at the National Spelling Bee. *Social Psychological & Personality Science, 2,* 174–181. http://dx.doi.org/10.1177/1948550610385872

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92,* 1087–1101. http://dx.doi.org/10.1037/0022-3514.92.6.1087

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (grit-s). *Journal of Personality Assessment, 91,* 166–174. http://dx.doi.org/10.1080/00223890802634290

Duckworth, A. L., Quinn, P. D., & Tsukayama, E. (2012). What No Child Left Behind leaves behind: The roles of IQ and self-control in predicting standardized achievement test scores and report card grades. *Journal of Educational Psychology, 104,* 439–451. http://dx.doi.org/10.1037/a0026280

Enders, C. K. (2010). *Applied missing data analysis.* New York, NY: Guilford Press.

Eskreis-Winkler, L., Shulman, E. P., Beal, S. A., & Duckworth, A. L. (2014). The grit effect: Predicting retention in the military, the workplace, school and marriage. *Frontiers in Psychology, 5,* 36. http://dx.doi.org/10.3389/fpsyg.2014.00036

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psycholog-

ical research. *Psychological Methods, 4,* 272–299. http://dx.doi.org/10.1037/1082-989X.4.3.272

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods, 14,* 275–299. http://dx.doi.org/10.1037/a0015825

Foster, D. P., & George, E. I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics, 22,* 1947–1975. http://dx.doi.org/10.1214/aos/1176325766

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74,* 59–109. http://dx.doi.org/10.3102/00346543074001059

Geiser, S., & Santelices, M. V. (2007). Validity of high school grades in predicting student success beyond the freshman year: High school record vs. standardized tests as indicators of four-year college outcomes. Research & Occasional Paper Series: CSHE. 6.07. *Center for Studies in Higher Education.* Retrieved from http://eprints.cdlib.org/uc/item/7306z0zf#page-1

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big-Five personality domains. *Journal of Research in Personality, 37,* 504–528. http://dx.doi.org/10.1016/S0092-6566(03)00046-1

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21,* 347–360. http://dx.doi.org/10.1111/j.1745-3984.1984.tb01039.x

Guo, B., Aveyard, P., Fielding, A., & Sutton, S. (2008). Testing the convergent and discriminant validity of the Decisional Balance Scale of the Transtheoretical Model using the Multi-Trait Multi-Method approach. *Psychology of Addictive Behaviors, 22,* 288–294. http://dx.doi.org/10.1037/0893-164X.22.2.288

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75,* 209–227. http://dx.doi.org/10.1007/s11336-010-9158-4

Harter, S. (2006). The self. In W. Damon (Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (6th ed., pp. 505–570). Hoboken, NJ: Wiley.

Hidi, S., & Renninger, A. (2006). The four-phase model of interest development. *Educational Psychologist, 41,* 111–127. http://dx.doi.org/10.1207/s15326985ep4102_4

Hoffman, J. L., & Lowitzki, K. E. (2005). Predicting college success with high school grades and test scores: Limitations for minority students. *The Review of Higher Education, 28,* 455–474. http://dx.doi.org/10.1353/rhe.2005.0042

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2,* 41–54. http://dx.doi.org/10.1007/BF02287965

Husman, J., & Lens, W. (1999). The role of the future in student motivation. *Educational Psychologist, 34,* 113–125. http://dx.doi.org/10.1207/s15326985ep3402_4

Ivcevic, Z., & Brackett, M. (2014). Predicting school success: Comparing conscientiousness, grit, and emotion regulation ability. *Journal of Research in Personality, 52,* 29–36. http://dx.doi.org/10.1016/j.jrp.2014.06.005

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38,* 1217–1218. http://dx.doi.org/10.1111/j.1365-2929.2004.02012.x

Joe, H., & Maydeu-Olivares, A. (2010). A general family of limited information goodness-of-fit statistics for multinomial data. *Psychometrika, 75,* 393–419. http://dx.doi.org/10.1007/s11336-010-9165-5

John, O., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). New York, NY: Guilford Press.

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V - Analysis of linear structural relationships by maximum likelihood and leas squares methods.* Chicago, IL: National Educational Resources.

Lens, W. (1986). Future time perspective: A cognitive-motivational concept. In D. R. Brown & J. Veroff (Eds.), *Frontiers of motivational psychology* (pp. 173–190). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4684-6341-5_10

Linn, M. C., & Eylon, B. (2006). Science education: Integrating views of learning and instruction. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 511–544). New York, NY: Routledge.

Lowney, P. (2013). *The effects of grit, conscientiousness, and question order of unsolvable problems on subsequent performance* (Unpublished thesis). Dublin, Ireland: DBS School of Arts. Retrieved from http://esource.dbs.ie/bitstream/handle/10788/1654/hdip_lowney_p_2013.pdf?sequence=1

Maddi, S. R., Matthews, M. D., Kelly, D. R., Villarreal, B., & White, M. (2012). The role of hardiness and grit in predicting performance and retention of USMA cadets. *Military Psychology, 24,* 19–28. http://dx.doi.org/10.1080/08995605.2012.639672

Maloney, E. A. (2016). Math anxiety: Causes, consequences, and remediation. In K. R. Wentzel & D. B. Miele (Eds.), *Handbook of motivation at school* (pp. 408–423). New York, NY: Taylor & Francis.

Marsh, H. W., Craven, R. G., Hinkley, J. W., & Debus, R. L. (2003). Evaluation of the big-two factor of academic motivation orientations: An evaluation of jingle-jangle fallacies. *Multivariate Behavioral Research, 38,* 189–224. http://dx.doi.org/10.1207/S15327906MBR3802_3

Marsh, H. W., & Hocevar, D. (1983). Confirmatory factor analysis of multitrait-multimethod matrices. *Journal of Educational Measurement, 20,* 231–248. http://dx.doi.org/10.1111/j.1745-3984.1983.tb00202.x

Marsh, H. W., Martin, A. M., Yeung, A. S., & Craven, R. G. (in press). Competence self-perceptions: A cornerstone of achievement motivation and the positive psychology movement. In A. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (2nd ed.). New York, NY: Guilford Press.

Maszk, P., Eisenberg, N., & Guthrie, I. K. (1999). Relations of children's social status to their emotionality and regulation: A short-term longitudinal study. *Merrill-Palmer Quarterly, 45,* 468–492.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of fit testing in multidimensional contingency tables. *Psychometrika, 71,* 713–732. http://dx.doi.org/10.1007/s11336-005-1295-9

Mischel, W., Cantor, N., & Feldman, S. (1996). Principles of self-regulation: The nature of willpower and self-control. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 329–360). New York, NY: Guilford Press.

Murphy, P. K., & Alexander, P. A. (2000). A motivated look at motivational terminology. *Contemporary Educational Psychology, 25,* 3–53. http://dx.doi.org/10.1006/ceps.1999.1019

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50–64. http://dx.doi.org/10.1177/01466216000241003

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66,* 543–578. http://dx.doi.org/10.3102/00346543066004543

Peluso, T., Ricciardelli, L. A., & Williams, R. J. (1999). Self-control in relation to problem drinking and symptoms of disordered eating. *Addictive Behaviors, 24,* 439–442. http://dx.doi.org/10.1016/S0306-4603(98)00056-2

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82,* 33–40. http://dx.doi.org/10.1037/0022-0663.82.1.33

Pintrich, P. R., & Schrauben, B. (1992). Students' motivational beliefs and their cognitive engagement in classroom academic tasks. *Student Perceptions in the Classroom, 7,* 149–183.

Pintrich, P. R., Smith, D. A., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ).* Ann Arbor, MI: National Center for Research to Improve Postsecondary Teaching and Learning; Retrieved from http://files.eric.ed.gov/fulltext/ED338122.pdf

Pintrich, P. R., & Zusho, A. (2007). Student motivation and self-regulated learning in the college classroom. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 731–810). Springer Netherlands. http://dx.doi.org/10.1007/1-4020-5742-3_16

Reschly, A. L., & Christenson, S. L. (2012). *Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct* (pp. 3–19). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-2018-7_1

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17,* 354–373. http://dx.doi.org/10.1037/a0029315

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement, 47,* 361–372. http://dx.doi.org/10.1111/j.1745-3984.2010.00118.x

Rojas, J. P., Reser, J. A., Usher, E. L., & Toland, M. D. (2012). *Psychometric properties of the academic grit scale.* Lexington, KY: University of Kentucky. Retrieved from http://sites.education.uky.edu/motivation/files/2013/08/PojasPeserTolandUsher.pdf

Rojas, J. P., & Usher, E. L. (2013). *Exploring correlations among creativity, grit, and mathematics achievement in socioeconomically diverse schools.* Lexington, KY: University of Kentucky. Retrieved from http://sites.education.uky.edu/motivation/files/2013/08/RojasUsher.pdf

Roth, P. L., BeVier, C. A., Switzer, F. S. I. I. I., & Schippmann, J. S. (1996). Meta-analyzing the relationship between grades and job performance. *Journal of Applied Psychology, 81,* 548–556. http://dx.doi.org/10.1037/0021-9010.81.5.548

Roth, P. L., & Clarke, R. L. (1998). Meta-analyzing the relation between grades and salary. *Journal of Vocational Behavior, 53,* 386–400. http://dx.doi.org/10.1006/jvbe.1997.1621

Samejima, F. (1972). A general model for free-response data. *Psychometrika, 37,* 1 Pt. 2, 1–68. Retrievd from http://eric.ed.gov/?id=EJ056490

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8,* 3–15. http://dx.doi.org/10.1191/096228099671525676

Schmid, J., & Leiman, J. (1957). The development of hierarchical factor solutions. *Psychometrika, 22,* 53–61. http://dx.doi.org/10.1007/BF02289209

Sehgal, P. (2015, December 1). *The profound emptiness of resilience* [Web article]. Retrieved from http://www.nytimes.com/2015/12/06/magazine/the-profound-emptiness-of-resilience.html?smid=fb-nytimes&smtyp=cur&_r=1

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika, 52,* 591–611. http://dx.doi.org/10.1093/biomet/52.3-4.591

Shoda, Y., Mischel, W., & Peake, P. K. (1990). Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification: Identifying diagnostic conditions. *Developmental Psychology, 26,* 978–986. http://dx.doi.org/10.1037/0012-1649.26.6.978

Skinner, E., Furrer, C., Marchand, G., & Kindermann, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic? *Journal of Educational Psychology, 100,* 765–781. http://dx.doi.org/10.1037/a0012840

Skinner, E. A., Kindermann, T. A., Connell, J. P., & Wellborn, J. G. (2009). Organizational constructs in the dynamics of motivational development. In K. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 223–245). Mahwah, NJ: Lawrence Erlbaum.

Skinner, E. A., Kindermann, T. A., & Furrer, C. J. (2008). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and Psychological Measurement, 69,* 493–525. http://dx.doi.org/10.1177/0013164408323233

Skinner, E. A., Wellborn, J. G., & Connell, J. P. (1990). What it takes to do well in school and whether I've got it: A process model of perceived control and children's engagement and achievement in school. *Journal of Educational Psychology, 82,* 22–32. http://dx.doi.org/10.1037/0022-0663.82.1.22

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika, 66,* 563–576. http://dx.doi.org/10.1007/BF02296196

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics, 19,* 279–281. http://dx.doi.org/10.1214/aoms/1177730256

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *The Journal of Consumer Research, 25,* 78–107. http://dx.doi.org/10.1086/209528

Strayhorn, T. L. (2014). What role does grit play in the academic success of Black male collegians at predominantly White institutions? *Journal of African American Studies, 18,* 1–10. http://dx.doi.org/10.1007/s12111-012-9243-0

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *Journal of Personality, 72,* 271–324. http://dx.doi.org/10.1111/j.0022-3506.2004.00263.x

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring.* Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., & Wainer, H. (2001). *Test scoring.* Hillsdale, NJ: Lawrence Erlbaum.

Thorsen, C., & Cliffordson, C. (2012). Teachers' grade assignment and the predictive validity of criterion-referenced grades. *Educational Research and Evaluation, 18,* 153–172. http://dx.doi.org/10.1080/13803611.2012.659929

Tough, P. (2012). *How children succeed: Grit, curiosity, and the hidden power of character.* New York, NY: Houghton Mifflin Harcourt.

Turiel, E., Chung, E., & Carr, J. A. (in press). Struggles for equal rights and social justice as unrepresented and represented in psychological research. In S. Horn, M. Ruck, & L. Liben (Eds.), *Vol. 1: Equity and justice in developmental sciences: Theoretical and methodological issues. Advances in Child Development and Behavior.* New York, NY: Elsevier. http://dx.doi.org/10.1016/bs.acdb.2015.11.004

U.S. Department of Education. (2013). *Promoting grit, tenacity, and perseverance: Critical factors for success in the 21st century.* Washington, DC: Office of Educational Technology. Retrieved from http://pgbovine.net/OET-Draft-Grit-Report-2-17-13.pdf

Wang, M. T., & Eccles, J. S. (2012). Social support matters: Longitudinal effects of social support on three dimensions of school engagement from middle to high school. *Child Development, 83,* 877–895. http://dx.doi.org/10.1111/j.1467-8624.2012.01745.x

West, M. R., Kraft, M. A., Finn, A. S., Martin, R., Duckworth, A. L., Gabrieli, C. F., & Gabrieli, J. D. (2016). Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis, 38,* 148–170. http://dx.doi.org/10.3102/0162373715597298

Whiteside, S. P., & Lynam, D. R. (2001). The five factor model and impulsivity: Using a structural model of personality to understand im-

pulsivity. *Personality and Individual Differences, 30,* 669–689. http://dx.doi.org/10.1016/S0191-8869(00)00064-7

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait–multimethod data. *Applied Psychological Measurement, 9,* 1–26. http://dx.doi.org/10.1177/014662168500900101

Wigfield, A., Eccles, J. S., Fredericks, J., Roeser, R., Schiefele, U., & Simpkins, S. (2015). Development of achievement motivation and engagement. In R. Lerner (Series Ed.) & C. Garcia Coll & M. Lamb (Volume Eds.), *Handbook of child psychology, 7th ed.Vol. 3, Social and emotional development.* New York, NY: Wiley. http://dx.doi.org/10.1002/9781118963418.childpsy316

Wigfield, A., Tonks, S., & Klauda, S. L. (in press). Expectancy-value theory. In K. R. Wentzel & D. Miele (Eds.), *Handbook of motivation in school* (2nd ed.). New York, NY: Routledge.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12,* 58–79. http://dx.doi.org/10.1037/1082-989X.12.1.58

Wolters, C. A. (2004). Advancing achievement goal theory: Using goal structures and goal orientations to predict students' motivation, cognition, and achievement. *Journal of Educational Psychology, 96,* 236–250. http://dx.doi.org/10.1037/0022-0663.96.2.236

Wolters, C. A., & Hussain, M. (2015). Investigating grit and its relations with college students' self-regulated learning and academic achievement. *Metacognition and Learning, 10,* 293–311. http://dx.doi.org/10.1007/s11409-014-9128-9

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika, 92,* 937–950. http://dx.doi.org/10.1093/biomet/92.4.937

Zaleski, Z. (1987). Behavioral effects of self-set goals for different time ranges. *International Journal of Psychology, 22,* 17–38. http://dx.doi.org/10.1080/00207598708246765

Zimmerman, B. J. (2011). Motivational sources and outcomes of self-regulated learning and performance. In B. J. Zimmerman & D. H. Schunk (Eds.), *Handbook of self-regulation of learning and performance* (pp. 49–64). New York, NY: Routledge.

Zimmerman, B. J., & Cleary, T. J. (2009). Motives to self-regulate learning: A social cognitive account. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 247–264). New York, NY: Routledge.

Zimmerman, B. J., & Schunk, D. H. (2011). *Handbook of self-regulation of learning and performance.* New York, NY: Routledge.

Zusho, A., & Edwards, K. (2011). Self-regulation and achievement goals in the college classroom. *New Directions for Teaching and Learning, 2011,* 21–31. http://dx.doi.org/10.1002/tl.441

Received February 1, 2016
Revision received August 16, 2016
Accepted August 22, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 5, 621–634

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000163

Math Self-Concept, Grades, and Achievement Test Scores: Long-Term Reciprocal Effects Across Five Waves and Three Achievement Tracks

A. Katrin Arens
German Institute for International Educational Research, Germany

Herbert W. Marsh
Australian Catholic University and Oxford University

Reinhard Pekrun
University of Munich and Australian Catholic University

Stephanie Lichtenfeld
University of Munich

Kou Murayama
University of Reading and Kochi University of Technology

Rudolf vom Hofe
University of Bielefeld

This study examines reciprocal effects between self-concept and achievement by considering a long time span covering grades 5 through 9. Extending previous research on the reciprocal effects model (REM), this study tests (1) the assumption of developmental equilibrium as time-invariant cross-lagged paths from self-concept to achievement and from achievement to self-concept, (2) the generalizability of reciprocal relations when using school grades and standardized achievement test scores as achievement indicators, and (3) the invariance of findings across secondary school achievement tracks. Math self-concept, school grades in math, and math achievement test scores were measured once each school year with a representative sample of 3,425 German students. Students' gender, IQ, and socioeconomic status (SES) were controlled in all analyses. The findings supported the assumption of developmental equilibrium for reciprocal effects between self-concept and achievement across time. The pattern of results was found to be invariant across students attending different achievement tracks and could be replicated when using school grades and achievement test scores in separate and in combined models. The findings of this study thus underscore the generalizability and robustness of the REM.

Keywords: math self-concept, math achievement, reciprocal effects, school tracks

Supplemental materials: http://dx.doi.org/10.1037/edu0000163.supp

Academic self-concept is defined as students' self-perceptions of competence in academic domains (Shavelson, Hubner, & Stanton, 1976). It has been a prominent construct in educational psychology over the last several decades as it has been found to share substantial relations to outcome variables including academic achievement (Marsh, 2007; Marsh & O'Mara, 2008b; Valentine, DuBois, & Cooper, 2004). In this context, many studies have supported reciprocal relations between academic self-concept and achievement involving important theoretical implications for self-concept formation and practical implications for the enhancement of both self-concept and achievement (for an overview see Marsh & Craven, 2006). However, several issues remain to be clarified. These include the assumption of developmental equilibrium, the interplay of school grades and standardized achievement test scores as two alternative achievement measures, and the generalizability of findings

This article was published Online First December 1, 2016.

A. Katrin Arens, Department of Education and Human Development, and Center for Research on Individual Development and Adaptive Education of Children (IDeA), German Institute for International Educational Research, Germany; Herbert W. Marsh, Institute of Positive Psychology and Education, Australian Catholic University, and Department of Education, Oxford University; Reinhard Pekrun, Department of Psychology, University of Munich, and Institute for Positive Psychology and Education, Australian Catholic University; Stephanie Lichtenfeld, Department of Psychology, University of Munich; Kou Murayama, School of Psychology and Clinical Language Sciences, University of Reading, and Research Institute, Kochi University of Technology; Rudolf vom Hofe, Faculty of Mathematics, University of Bielefeld.

This research was supported by four grants from the German Research Foundation to R. Pekrun (PE 320/11-1, PE 320/11-2, PE 320/11-3, PE 320/11-4). We thank the German Data Processing and Research Center of the International Association for the Evaluation of Educational Achievement for organizing the sampling and performing the assessments.

Correspondence concerning this article should be addressed to A. Katrin Arens, Department of Education and Human Development, and Center for Research on Individual Development and Adaptive Education of Children (IDeA), German Institute for International Educational Research, Schloßstraße 29, D-60486 Frankfurt am Main, Germany. E-mail: arens@dipf.de

across school tracks. These issues are targeted in the present study.

Relations Between Academic Self-Concept and Achievement

When examining the link between academic self-concept and academic achievement, many studies have attested substantial cross-sectional relations (e.g., Arens, Yeung, Craven, & Hasselhorn, 2011; Marsh et al., 2013). Studies scrutinizing longitudinal relations have attracted even more attention because they help elucidate causality in the relation between self-concept and achievement. Thus, a critical question has been whether self-concept is an outcome of achievement or whether achievement is an outcome of self-concept. Calsyn and Kenny (1977) posited two models for the temporal relation between self-concept and achievement. While the skill development model suggests that achievement predicts self-concept, the self-enhancement model suggests that self-concept predicts achievement. Originally, the skill development and self-enhancement models were strictly contrasted but recent research indicates that such a clear either-or stance is inappropriate because self-concept and achievement share mutually reinforcing relations. Therefore, in contemporary self-concept research, the reciprocal effects model (REM) prevails for depicting the relations between self-concept and achievement. Accordingly, self-concept is both an outcome of former and a predictor of subsequent achievement (e.g., Huang, 2011; Marsh & Craven, 2006; Möller, Retelsdorf, Köller, & Marsh, 2011; Niepel, Brunner, & Preckel, 2014).

Number of Waves and Developmental Equilibrium

Studies integrating two measurement waves can already serve to test the temporal ordering of relations between self-concept and achievement (Marsh, Byrne, & Yeung, 1999). However, the inclusion of three or more waves allow for the examination and comparison among skill development and self-enhancement effects over time. The assumption of developmental equilibrium would expect skill development and self-enhancement paths to be of similar size from one wave to the next (for studies integrating related assumptions see, e.g., Marsh, Craven, et al., 2016; Marshall, Parker, Ciarrochi, & Heaven, 2014). Hence, in this case the effect from achievement (self-concept) to self-concept (achievement) would be of similar size across the different time lags, for example, across Waves 1 and 2 and Waves 2 and 3.

Developmental equilibrium is not essential to providing evidence of reciprocal effects, but support for this assumption has a number of important advantages. For complex models resulting from the assessment of self-concept and achievement across many waves with many items used in each wave, the added parsimony provides more robust and precise estimates and facilitates the presentation and interpretation of results. Moreover, support for developmental equilibrium offers some protection against alternative interpretations of the results based on potential other variables not considered (Kenny, 1975). More importantly, if developmental equilibrium can be supported, self-concept (achievement) exerts a similar influence on later achievement (self-concept) at different time points. Hence, studies testing developmental equilibrium are best based on a large number of measurement waves which cover a long and relevant period of time. For instance, it would be interesting to examine the robustness of skill development and self-enhancement effects across adolescence or secondary school years.

However, in a meta-analysis, Huang (2011) revealed that out of 32 studies examining longitudinal relations between self-concept and achievement, 19 studies relied on a two-wave design, eight studies had three measurement waves, two studies were respectively based on four and five measurement waves, and only one study covered six measurement waves. Thus, there seems to be a need for further studies that include more than two or three measurement waves. In line with these considerations, the present study covers five waves tracking German students from the fifth to the ninth grade. Students' self-concept and achievement were collected once every school year. Therefore, the present study can replicate findings on the REM over an exceptionally long time interval and examine developmental equilibrium across students' fifth to ninth grade, the years of mandatory secondary schooling in Germany.

First-Order and Higher Order Paths

The REM is commonly tested by cross-lagged panel models embedded in the framework of structural equation modeling (SEM; Curran & Bollen, 2001; Marsh et al., 1999). This modeling approach includes autoregressive or stability paths estimating the effect of one variable on the same variable across subsequent measurement waves, for example, the relation between self-concept measured at Time 1 (T1) and self-concept measured at Time 2 (T2). In addition, cross-lagged paths represent the reciprocal relations of one variable on another variable between measurement waves (i.e., effects of self-concept at T1 on achievement at T2 and effects of achievement at T1 on self-concept at T2). In studies that cover more than two measurement waves, it is possible to include both first-order and higher order paths. First-order paths depict the relation between two directly adjacent time points, that is, the effect of self-concept at T1 on self-concept at T2 as an example of a first-order stability path, and the effect of self-concept at T1 on achievement at T2 as an example of a first-order cross-lagged path. First-order paths thus describe "lag 1" paths referring to the effect of one variable on the same variable across two adjacent measurement waves. Higher-order paths describe the relations between constructs measured at more distal time points. Second-order paths refer to "lag 2" paths among constructs measured at T1 and T3, third-order paths refer to "lag 3" paths among constructs measured at T1 and T4 and so on. Thus, for instance, second-order stability addresses the relation between self-concept measured at T1 and self-concept measured at T3, and second-order cross-lagged paths depict the reciprocal relations between self-concept at T1 and achievement at T3.

Beyond the inherent inclusion of first-order paths, it might be worthwhile to consider higher order paths in cross-lagged panel models for the reciprocal relation between self-concept and achievement as this allows for examining long term relations (Marsh & O'Mara, 2008a). Indeed, in a four-wave model for studying the relations between reading self-concept and reading achievement, besides first-order stability paths, Retelsdorf, Köller, and Möller (2014) found significant higher order (i.e., second-order and third-order) stability estimates for reading self-concept

and reading achievement. In addition to the corresponding first-order path, there was also evidence of significant higher order cross-lagged paths for the relation between former reading achievement and later reading self-concept. Marsh, Gerlach, Trautwein, Lüdtke, and Brettschneider (2007) conducted a three-wave study examining the longitudinal relations between physical self-concept and physical achievement. The results revealed first-order and second-order stability estimates for both physical self-concept and physical performance. In addition, there was evidence of first-order and second-order cross-lagged paths between physical achievement and physical self-concept. Accordingly, the recommendations formulated by Marsh et al. (1999) for "ideal" studies on the REM include the advice to start with a full-forward model which incorporates the estimations of all paths. Researchers are then advised to compare this complete model with more parsimonious alternative models. Therefore, in this study, we use a full-forward model including second-order, third-order, and fourth-order paths as a starting point to examine reciprocal relations between self-concept and achievement across five measurement waves.

Achievement Indicators: School Grades Versus Test Scores

School grades and standardized achievement test scores are the two most commonly used indicators of students' achievement. School grades are very salient to students as they are directly communicated, easy to compare among classmates, and entail important implications for students' school careers. School grades do not only narrowly represent student achievement but also refer to other student characteristics such as students' effort or classroom behavior (Brookhart, 1993; McMillan, Myran, & Workman, 2002; Zimmermann, Schütte, Taskinen, & Köller, 2013). On the other hand, students are often unaware of their relative performance on standardized achievement tests. Therefore, students' self-concept has been found to be more strongly related to school grades than to standardized achievement test scores (Marsh et al., 2014; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005).

Nonetheless, school grades suffer from idiosyncrasies as teachers have been found to grade on a curve, allocating the best grades to the relatively best performing students within a classroom and the poorest grades to the relatively poorest performing students (Marsh et al., 2014). Hence, teachers use the classroom as a narrow frame of reference in their grading procedure. Accordingly, the same student with the same level of objective achievement can receive divergent grades depending on the average achievement and achievement standards in the individual student's class. For this reason, school grades are difficult to compare across classes, schools, and nations whereas standardized achievement tests are particularly designed for the purpose of such comparisons. Thus, school grades and standardized achievement test scores each have their advantages and disadvantages when they are used as achievement indicators, emphasizing their distinct yet complementary nature.

Therefore, it appears worthwhile to use both kinds of achievement indicators to examine reciprocal relations to self-concept. Nonetheless, the majority of studies supporting reciprocal relations between self-concept and achievement have included school grades as achievement indicators (e.g., Marsh, 1990; Niepel et al., 2014; for an overview see Huang, 2011). Yet, there is also evidence that reciprocal relations between self-concept and achievement exist when using standardized achievement test scores as achievement indicators (Möller, Zimmermann, & Köller, 2014; Retelsdorf et al., 2014; Seaton, Parker, Marsh, Craven, & Yeung, 2014). However, most previous studies have considered school grades or achievement test scores separately whereas a more sophisticated approach would be to consider both achievement indicators simultaneously. The study of Marsh, Trautwein, et al. (2005) provided evidence of the REM for math self-concept and math achievement when analyzing math grades and math test scores separately as well as when combining them into one model. As this study included only two measurement waves, there still seems to be a need for studies that combine both achievement indicators and consider multiple waves across a longer time interval to more adequately test the validity of the REM and the assumption of developmental equilibrium for both kinds of achievement indicators.

Generalizability Across School Tracks

A number of studies have documented the applicability of the REM to both academic and nonacademic domains such as reading (Retelsdorf et al., 2014), math (Marsh, Trautwein, et al., 2005), and physical ability (Marsh et al., 2007), indicating its generalizability across different content domains. The REM has also been tested regarding its generalizability across different student characteristics such as age (Guay, Marsh, & Boivin, 2003) and gender (Marsh, Trautwein, et al., 2005). Other studies further indicated the cross-cultural generalizability of the REM because reciprocal relations between self-concept and achievement have been found with Australian (Seaton et al., 2014), American (Marsh & O'Mara, 2008a), German (Marsh, Trautwein, et al., 2005), Hong Kong (Marsh, Hau, & Kong, 2002), and French Canadian (Guay et al., 2003) students.

To the best of our knowledge, research is lacking regarding the generalizability of the REM across students attending different school tracks. This is surprising because many educational systems implement at least some kind of tracking, particularly in secondary education (Chmielewski, Dumont, & Trautwein, 2013; LeTendre, Hofer, & Shimizu, 2003). Students attending different achievement tracks seem to differ in various respects. For instance, they have been found to reveal different levels of academic achievement, motivation (e.g., interest), and academic self-concept (e.g., Baumert, Watermann, & Schümer, 2003; Becker, Lüdtke, Trautwein, & Baumert, 2006; Hanushek & Wößmann, 2006; Köller & Baumert, 2001). This finding might be partly due to differences in the students' learning environments. Students attending high-achievement tracks have been found to receive higher levels of instructional quality and to experience fewer disciplinary problems in the classroom, but also to get lower levels of individual learning support from the teacher compared to students in lower achievement track schools (Klieme & Rakoczy, 2003; Kunter et al., 2005). The present study aims to investigate and compare the REM among students attending different secondary school tracks as this provides an opportunity to test the generalizability of the REM across students experiencing different learning environments and educational opportunities.

Controlling for Covariates

The REM posits achievement to be a major determinant of self-concept and self-concept to be a major determinant of achievement. However, students' socioeconomic status (SES), IQ, and gender are also known to affect students' achievement as well as students' self-concept. Hence, studies aiming to establish reciprocal relations between self-concept and achievement would do well to consider these background variables.

Students from lower SES families demonstrate lower levels of achievement (Bradley & Corwyn, 2002; Sirin, 2005). High levels of student achievement have also often been linked to a high IQ (Frey & Detterman, 2004; Furnham & Monsen, 2009; Spinath, Spinath, & Plomin, 2008). Furthermore, student achievement is associated with gender. Specifically, girls display higher achievement in verbal subjects (De Fraine, Van Damme, & Onghena, 2007; Van de gaer, Pustjens, Van Damme, & De Munter, 2008) including reading (Lietz, 2006; Mullis, Martin, Kennedy, & Foy, 2007). The findings related to math are less clear and seem to vary contingent upon the achievement indicator. Some studies have found higher scores for boys on standardized math achievement tests (Brunner, Krauss, & Kunter, 2008; Matteucci & Mignani, 2011; Van de gaer et al., 2008), but other studies have reported no or only small gender differences (Hyde, Fennema, & Lamon, 1990; Nowell & Hedges, 1998). When considering school grades in math, girls were found to obtain higher grades than boys (Marsh & Yeung, 1998), although other studies could not find any gender differences (Marsh, Trautwein, et al., 2005). Regarding domain-specific academic self-concepts, there is consistent evidence for gender differences which are in line with gender stereotypes. Hence, girls show higher levels of verbal self-concept whereas boys display higher levels of math self-concept (Fredricks & Eccles, 2002; Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002; Skaalvik & Skaalvik, 2004).

The Present Study

Using a large representative sample of German secondary school students, the present study examines reciprocal relations between math self-concept and math achievement. As an innovative contribution to existing research, the study covers a long time span with five measurement waves enabling a proper investigation of the assumption of developmental equilibrium across German students' mandatory secondary school years. Moreover, this study examines the generalizability of reciprocal effects between math self-concept and math achievement across German students attending three different achievement tracks. The analytic approach starts with a complex full-forward model before turning to more parsimonious models in order to consider and test the adequacy of first-order and higher order stability and cross-lagged paths. School grades and standardized achievement test scores are simultaneously considered in combined models to account for the two most widely used yet distinctive achievement indicators (Marsh, Trautwein, et al., 2005; Marsh et al., 2014). Finally, gender, IQ, and SES are considered as covariates to control for other important variables influencing students' self-concept and achievement.

Method

Sample

The data analyzed in this study originate from the Project for the Analysis of Learning and Achievement in Mathematics (PALMA; Frenzel, Goetz, Lüdtke, Pekrun, & Sutton, 2009; Frenzel, Pekrun, Dicke, & Goetz, 2012; Marsh, Pekrun, Lichtenfeld et al., 2016; Marsh, Pekrun, Parker, et al., 2016; Murayama, Pekrun, Lichtenfeld, & vom Hofe, 2013; Murayama, Pekrun, Suzuki, Marsh, & Lichtenfeld, 2016; Pekrun, Lichtenfeld, Marsh, Murayama, & Goetz, in press). PALMA is a large-scale longitudinal study investigating the development of math achievement and its determinants (e.g., math-related motivation, classroom instruction, family variables) during secondary school in Germany. The study was conducted in the German federal state of Bavaria and covers six measurement waves spanning grades 5 to 10 with one measurement point each school year. Sampling and the assessments were conducted by the German Data Processing and Research Center of the International Association for the Evaluation of Educational Achievement (IEA). The sample represented the typical student population in the German federal state of Bavaria in terms of secondary school achievement tracks and student characteristics such as gender, urban versus rural location, and SES. Participation rate at the school level was 100%. At the first measurement wave in grade 5, the sample comprised 2,070 students (49.6% female, 37.2% low-achievement track students, 27.1% middle-achievement track students, and 35.7% high-achievement track students). The students then had a mean age of 11.75 years (SD = 0.68), which is the typical age for fifth grade students in Germany. A number of 42 schools participated in the study and two classes were randomly drawn within each school for final participation. For the subsequent data collections, the study did not only track the students who had already participated in the earlier assessments, but also included students who more recently entered classrooms participating in the PALMA study and thus had not yet participated in the study (for more details on the sampling procedure, see Pekrun et al., 2007).

In the German federal state of Bavaria, beginning in grade 5, students are allocated to either low-achievement (Hauptschule), middle-achievement (Realschule), or high-achievement (Gymnasium) tracks. This decision is mainly based on students' achievement in the fourth grade of elementary school. Low-achievement track students commonly leave school after the ninth grade with a qualification allowing them to apply for an apprenticeship, middle-achievement track schools end after the tenth grade and students may begin vocational training, and high-achievement track students attend school until the thirteenth grade after which they may enter university. For reasons of including low-achievement track students, the present study focuses on the first five measurement waves covering students grade 5 to 9. The final sample of the present study consists of 3,425 students (n = 1,714, 50.0% girls; n = 1,710, 49.9% boys; n = 1, 0.01% no gender) and included all students who participated in at least one of the five assessments. Among this final sample, 1,187 students attended the high-achievement track, 1,050 the middle-achievement track, and 1,188 the low-achievement track. Of the final sample, 38.7% participated in all five measurement waves (i.e., grades 5 to 9), and 9.0%, 18,9%, 15.1%, and 18.3% took part in four, three, two, or one of the assessments, respectively.

The students answered a questionnaire toward the end of each successive school year. All instruments were administered in the students' classrooms by trained external test administrators. Participation in the study was voluntary and parental consent was obtained for every student. Each survey was depersonalized to ensure participant confidentiality.

Measures

Math self-concept. Math self-concept was measured by the PALMA six-item math self-concept scale at each measurement wave. The items (i.e., "In math, I am a talented student"; "It is easy to understand things in math"; "I can solve math problems well"; "It is easy for me to write math tests"; "It is easy for me to learn something in math"; "If the math teacher asks a question, I can answer it correctly most of the time") were answered using a 5-point Likert scale (1 = *not at all true* to 5 = *completely true*). The scale showed high reliability at each of the five measurement waves both when using the coefficient alpha reliability estimate (α) and when using the scale reliability estimate (ρ; also labeled as *composite* or *instrument reliability*) which was explicitly established within the framework of SEM (Raykov, 2009; T1: $\alpha = .876$, $\rho = .879$; T2: $\alpha = .895$, $\rho = .896$; T3: $\alpha = .893$, $\rho = .894$; T4: $\alpha = .910$, $\rho = .910$; T5: $\alpha = .920$, $\rho = .921$).

Math achievement. Students' math achievement was measured both in terms of school grades and standardized achievement test scores. Math school grades were retrieved from school documents in terms of the report cards students received at the end of each school year, reflecting students' average math accomplishments throughout the school year. In Germany, school grades range from 1 (*highest achievement*) to 6 (*lowest achievement*). For ease of interpretation, the grades were recoded prior to all analyses so that higher grades represent higher achievement.

The Regensburg Mathematical Achievement Test (vom Hofe, Kleine, Blum, & Pekrun, 2005; vom Hofe, Pekrun, Kleine, & Götz, 2002) was used to assess students' standardized math achievement test scores. This test was explicitly designed for the PALMA study serving to measure students' development of math competencies across secondary school years. Thus, different test versions are available for the different grade levels. The conception of this test has substantially and methodologically been linked to the concept underlying the math tests applied in the Programme for International Student Assessment (PISA). Following the construct of mathematical literacy, the test operationalizes math competences as mathematical modeling and problem solving in terms of students' abilities to convert real-world problems into mathematical models, to solve these problems in the context of mathematical models, and to transfer the solutions to reality. On the basis of this conceptualization, the test targets students' modeling competencies and algorithmic competencies in arithmetic, algebra, and geometry. Methodologically, the test was constructed within the framework of item response theory (IRT; Wu, Adams, Wilson, & Haldane, 2007). At each measurement point, students worked on one of two different parallel test versions with 60 to 90 items each, the exact number of items varying across waves. The items were formulated either as multiple-choice or open-ended items. Prespecified guidelines were given to two trained raters to score the open-ended items. The ratings showed a high level of interrater agreement supporting their objectivity (i.e., interrater disagreement for the test version A [the parallel to version B] was 0.04% [0.13%], thus 0.085% on average). Depending on the measurement wave, approximately 20 anchor items served to link the two parallel test versions within each measurement point and the different tests across the five measurement points. Achievement test scores were scaled using one-parameter logistic IRT applying concurrent calibration (Rasch scaling; Wu et al., 2007) that has been found to have many advantages (e.g., model parsimony, parameter linearity) relative to alternative models (Liu, 2010; Wright, 1999) and that was also used in previous studies utilizing the PALMA math achievement test (Murayama et al., 2013). The reliability of item separation in IRT scaling was 0.99.

Covariates. Students' gender, IQ, and SES measured at T1 served as covariates. Students' IQ was measured using the German adaptation of Thorndike's Cognitive Abilities Test (Kognitiver Fähigkeitstest, KFT 4–12 + R; Heller & Perleth, 2000). Reliability of the 25 item scale was $\alpha = .934$. Supporting validity, the IQ test scores were found to be substantially correlated with students' math achievement test scores (T1: $r = .577$, $p < .01$) and to discriminate between students of the different achievement tracks (T1: $F(2, 1987) = 327.778$, $p < .001$), with students attending the high-achievement track displaying the highest mean levels ($M = 110.03$, $SD = 10.36$), followed by the middle-achievement track students ($M = 104.535$, $SD = 9.47$). Students from the low-achievement track displayed the lowest IQ mean level ($M = 96.20$, $SD = 11.91$).

SES was assessed by parental report using the Erikson Goldthorpe Portocarero (EGP) social class scheme (Erikson, Goldthorpe, & Portocarero, 1979). The EGP consists of six ordered categories of parental occupational status wherein higher values represent higher SES.

Statistical Analyses

The analyses were conducted within the SEM framework with *Mplus* 7.5 (Muthén & Muthén, 1998-2015). All models were conducted using the robust maximum likelihood (MLR) estimator which is robust against non-normality of the observed variables (Hox, Maas, & Brinkhuis, 2010; Muthén & Muthén, 1998-2015). The *Mplus* option < type = complex > was used to accommodate the hierarchical nature of the study. Specifically, students were nested within the 42 participating schools and students attending the same school might be more similar to each other than students from different schools, resulting in nonindependence of observations. Failure to attend to the hierarchical nature of the data could lead to biased standard errors—a miscalculation that is corrected by this *Mplus* model command (Muthén & Satorra, 1995).[1]

As inherent in any longitudinal study, the data set consisted of missing values on the measured variables which should be appropriately dealt with. In this study, missingness on variables mainly originates from the fact that students entered the study at later measurement waves without having completed the measures at earlier waves. The attrition rate could be kept rather low during the time period covered in the present study, that is, up to grade 9 after which the attrition rate is higher due to the low-achievement track students' leaving school. More concretely, among the total sample

[1] It was not possible to use students' classes as a clustering variable because the composition of students' classes changed across time.

of the study across all measurement waves, 60.4%, 60.1%, 69.9%, 70.3%, and 73.6% participated in the first, second, third, fourth, and fifth measurement wave respectively. Within each wave, the number of missing values was low for the self-concept measures (T1: 0.68% to 1.74%; T2: 0.29% to 1.70%; T3: 0.50% to 1.67%; T4: 0.46% to 1.78%; T5: 0.63% to 1.31%), the math achievement test scores (0.00% to 0.28%), and school grades in math at T1 to T4 (0.00% to 2.39%). Missing values on these variables were handled by the full information maximum likelihood estimator (FIML) implemented in *Mplus* by default (see Wang & Wang, 2012). FIML has been found to result in trustworthy, unbiased estimates for missing values (Enders, 2010; Graham, 2009) and represents an adequate means of managing missing data in longitudinal study designs (Jelicić, Phelps, & Lerner, 2009). However, *Mplus* excludes cases with missing data on any covariates if only defined as exogenous variables (Muthén & Muthén, 1998-2015). To use FIML for missing data on the continuous covariates (i.e., IQ and SES) as well, covariances among these covariates were estimated.

The amount of missing values was high (27.57%) regarding math school grades for the last measurement wave (T5) as the low-achievement track students left school after grade 9. Because of this high amount of missing data, we applied the technique of multiple imputation to handle missing data on students' math school grades for T5 which were imputed using the math school grades the students had obtained at the previous waves. Five sets of imputed data were created which were used for all analyses involving school grades and combined afterward (Little & Rubin, 2002) while retaining the FIML approach for estimating missing data on the remaining variables (i.e., self-concept at all waves, test scores at all waves, and school grades at T1 to T4).

In all models, one factor for math self-concept was assumed for each measurement wave defined by the six self-concept items answered by the students at the corresponding waves. Correlated uniquenesses for the same self-concept items over time were included in these models to account for the shared method variance due to the repeated use of the same items (Marsh & Hau, 1996). In addition, for each measurement wave, the models included two single-indicator achievement factors defined by students' school grades in math and their math achievement test scores, respectively.

The analyses started with a longitudinal confirmatory factor analysis model assuming separate self-concept and achievement factors for each of the five waves. In this model, the self-concept and achievement factors were freely estimated across time with the same set of items used to define the same number of factors at each measurement wave (configural invariance, Millsap, 2011). The analyses continued with a model of longitudinal measurement invariance by constraining the factor loadings to be of equal size across measurement waves (weak measurement invariance; Millsap, 2011). This model served to test whether the same constructs were measured at the different measurement waves (Widaman, Ferrer, & Conger, 2010).

To examine the generalizability of findings across students from different achievement tracks, students' attended school track (high-achievement, middle-achievement, or low-achievement track) was entered as a grouping variable in all models. To test whether the same constructs were measured in all groups of school tracks, we estimated a model assuming invariant factor loadings in the three groups of school tracks. We then stated an even more restrictive model by constraining the factor loadings to be simultaneously invariant across measurement waves and group to ensure that the same constructs were measured at each measurement wave in the three groups of students from different achievement tracks.

To test reciprocal relations between self-concept and achievement, we applied cross-lagged panel models and started with a full-forward model. The full-forward model included all possible (i.e., first-order and higher order) paths for the stability and the cross-lagged relations among the constructs, and additionally assumed the disturbances of constructs to be correlated within each wave (Figure 1; Marsh et al., 1999). Based on this complex model, we evaluated more parsimonious models with fewer paths. In this context, we first assessed the need to include both first-order and higher order paths by comparing the full-forward model with a model only including first-order stability and cross-lagged paths. Afterwards, we tested whether it is advantageous to incorporate first-order and higher order paths for both the stability and cross-lagged paths. For this purpose, we estimated a model including first-order and higher order stability paths but only first-order cross-lagged paths, and a model with first-order and higher order cross-lagged paths but only first-order stability paths. In the next step, the three covariates (gender, IQ, and SES) measured at T1

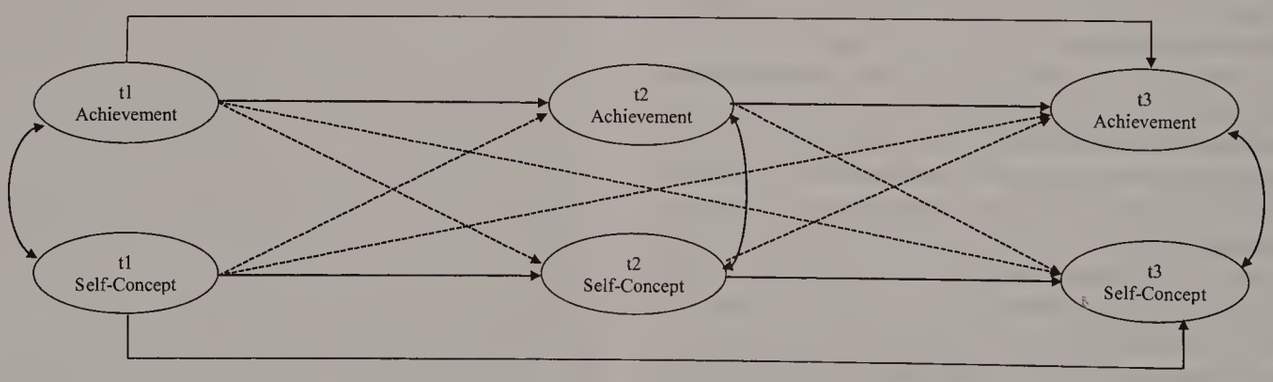


Figure 1. Prototype full-forward cross-lagged effects model for reciprocal relations between self-concept and achievement. For simplification, only three measurement waves are presented. Ovals represent latent constructs (self-concept and achievement factors); straight dashed arrows represent first-order and higher order (here: second-order) cross-lagged effects paths; straight solid arrows represent first-order and higher order (here: second-order) stability paths; curved arrows represent covariances between factors.

were included in the selected model to examine whether the findings were robust when controlling for these variables.

So far, the cross-lagged paths depicting the longitudinal effects between self-concept (achievement) and achievement (self-concept) were freely estimated across time. To test the assumption of developmental equilibrium, invariance constraints were imposed on these paths. The cross-lagged paths from one variable in one wave to another variable in a subsequent wave were assumed to be of equal size across all measurement points (e.g., achievement T1 → self-concept T2 = achievement T2 → self-concept T3 = achievement T3 → self-concept T4 = achievement T4 → self-concept T5). In a subsequent model, invariance constraints on the stability paths were included in terms that all stability estimates for one construct were restricted to be of the same size.

For presenting the relations between self-concept and achievement, we report the StdYX standardized coefficients provided by *Mplus* (Muthén & Muthén, 1998–2015), except for the effects of gender. The StdYX solution is based on the variances of both the continuous independent latent variable (X; e.g., math achievement) and the outcome (dependent) variable (Y; e.g., math self-concept) and interpreted as the mean change of Y in standard deviation units of Y for one standard deviation change in X. For gender as a binary variable, a proper standardized estimate results from standardizing the dependent variable Y only, which is provided by the StdY solution in Mplus and depicts the change in Y (e.g., math self-concept) in Y standard deviation units when X (i.e., gender) changes from zero to one.

To assess the fit of the models, we rely on a range of commonly applied descriptive goodness-of-fit indices (Marsh, Hau, & Grayson, 2005). We thus report the comparative fit index (CFI), the Tucker-Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardized root-mean-square residual (SRMR). Values above .90 and .95 for the CFI and TLI represent acceptable and good fit, respectively (Hu & Bentler, 1999). With respect to the RMSEA, values near .05 imply "close fit," values near .08 indicate "fair fit," and values above .10 represent "poor fit" (Browne & Cudeck, 1993). SRMR values below .05 are interpreted as a good model fit (Diamantopoulos & Siguaw, 2000) but Hu and Bentler (1999) proposed a less strict cut-off criterion of .08.

The invariance models can be conceptualized as nested models which only differ from each other in the parameters which were set invariant across time and group. To evaluate invariance, we follow the guidelines proposed by Cheung and Rensvold (2002) and Chen (2007), according to which invariance should not be rejected if $\Delta CFI \leq -.01$ and $\Delta RMSEA \leq +.015$ for the more restrictive as compared with the less restrictive model. The cut-off values suggested for the evaluation of latent models including invariance models should be considered as rough guidelines instead of golden rules. Researchers are rather advised to take all available information into account for an ultimate judgment of latent models, including parameter estimates, statistical conformity, and theoretical adequacy of the model besides the fit indices (Marsh, Hau, & Wen, 2004).

Results

The analyses reported herein are based on models including both school grades and standardized achievement test scores as achievement indicators, but the same series of models was also estimated using school grades and achievement test scores separately. The results from these models (i.e., using either school grades or achievement test scores) are reported in the online supplemental materials (see Models S1 to S7) and are fully consistent with the findings for the models combining school grades and achievement test scores.

The series of analyses started with a longitudinal measurement model assuming separate factors for math self-concept, math achievement test scores, and math school grades at each measurement wave (Model 1 in Table 1). The fit of this model was excellent and largely maintained when imposing invariance of factor loadings across time (Model 2) indicating that the same constructs were measured at each wave. Model 3 included students' achievement tracks as a grouping factor. The fit indices of this model still indicated good fit which was mostly retained when including invariant factor loadings across school tracks (Model 4). This finding indicates that the same constructs were measured in the three groups of students attending different achievement tracks. The fit indices also remained in the area of good model fit when assuming an even more restrictive model with invariant factor loadings both across measurement waves and school tracks (Model 5). This finding allowed for meaningful longitudinal analyses and comparisons across school tracks.

Model 6 is the full-forward model for describing the longitudinal relations between math self-concept, math achievement test scores, and math school grades for students from different achievement tracks. This model incorporates all possible first-order and higher order paths and only replaces the correlations among constructs by path coefficients. Therefore, it is statistically equivalent to Model 5 and results in the same fit. The path coefficients of Model 6 (see Table S6 in the online supplemental materials) suggest that multicollinearity might be at play since some coefficients for self-concept–achievement relations had implausible negative and small coefficients (Marsh, Dowson, Pietsch, & Walker, 2004). Hence, the full-forward model (Model 6) was compared to a less complex model (Model 7), which included only first-order stability and cross-lagged paths. The fit of Model 7 declined substantially compared to the full-forward Model 6 ($\Delta CFI = -.017$; $\Delta RMSEA = +.008$), suggesting that the inclusion of any higher order paths seems to be warranted. In Model 8, we integrated first-order and higher order stability paths, but only first-order cross-lagged paths. Here, the fit was highly similar to the fit resulting from the full-forward Model 6. However, when assuming first-order and higher order cross-lagged paths along with first-order stability paths (Model 9), the model fit substantially declined compared to the full-forward Model 6 ($\Delta CFI = -.015$; $\Delta RMSEA = +.008$). The findings thus argue for the integration of higher order stability paths but not for the inclusion of higher order cross-lagged paths leading us to keep Model 8.

The findings resulting from Model 8 were retained when adding the effects of the three covariates, that is, gender, IQ, and SES, which were assumed to be related to students' math self-concept, math achievement test scores and math school grades at T1 in Model 10. The covariates demonstrated significant effects on math self-concept, math school grades, and math achievement test scores (see Table S7 of the online supplemental materials).

Model 11 then served to test the assumption of developmental equilibrium. For this purpose, the first-order cross-lagged paths

Table 1
Goodness-of-Fit Indices of the Models Including School Grades and Test Scores as Achievement Indicators

Model	χ^2	*df*	CFI	TLI	RMSEA	SRMR	
1	1121.292	585	.990	.987	.016	.023	CFA longitudinal measurement model
2	1206.239	605	.989	.986	.017	.027	CFA longitudinal measurement model; invariance of factor loadings across time
3	2634.556	1755	.984	.978	.021	.029	Multi-group CFA longitudinal measurement model
4	2695.456	1805	.984	.979	.021	.030	Multi-group CFA longitudinal measurement model; invariance of factor loadings across school tracks
5	2780.448	1824	.982	.977	.021	.034	Multi-group CFA longitudinal measurement model; invariance of factor loadings across school tracks and time
6	2780.446	1824	.982	.977	.021	.034	Full-forward cross-lagged panel model; all paths freely estimated across school tracks and time
7	3857.313	1986	.965	.959	.029	.058	Cross-lagged panel model; only first-order stability and cross-lagged paths; all paths freely estimated across school tracks and time
8	2951.810	1932	.981	.977	.022	.036	Cross-lagged panel model; first-order and higher order stability paths, but only first-order cross-lagged paths; all paths freely estimated across school tracks and time
9	3641.541	1878	.967	.959	.029	.052	Cross-lagged panel model; first-order and higher order cross-lagged paths, but only first-order stability paths; all paths freely estimated across school tracks and time
10	3540.821	2265	.977	.972	.022	.037	Cross-lagged panel model; first-order and higher order stability paths, but only first-order cross-lagged paths; inclusion of control variables; all paths freely estimated across school tracks and time
11	3769.556	2331	.974	.970	.023	.042	Cross-lagged panel model; first-order and higher order stability paths, but only first-order cross-lagged paths; inclusion of control variables; invariance of cross-lagged paths across school tracks and time (developmental equilibrium)
12	3996.579	2409	.971	.968	.024	.052	Cross-lagged panel model; first-order and higher order stability paths, but only first-order cross-lagged paths; inclusion of control variables; invariance of cross-lagged paths and (first-order and higher order) stability paths across school tracks and time
13	4031.404	2427	.971	.968	.024	.052	Cross-lagged panel model; first-order and higher order stability paths, but only first-order cross-lagged paths; inclusion of control variables; invariance of cross-lagged paths, (first-order and higher order) stability, and covariates paths across school tracks and time

Note. All models are estimated with the robust maximum likelihood estimator; all chi-squares are significant ($p < .05$). CFA = confirmatory factor analyses; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

between former self-concept and later achievement (both school grades and test scores) and the first-order cross-lagged paths between former achievement and later self-concept were respectively set to be equal across all time lags and groups of school tracks (Model 11). The various goodness-of-fit indices of Model 11 still indicated good model fit and did not demonstrate a substantial decline (ΔCFI = $-.003$; ΔRMSEA = $+.001$) relative to the precedent less restrictive Model 10 so that the assumption of developmental equilibrium, which is additionally invariant across school tracks, could be supported.

Model 12 extends Model 11 in terms of invariance constraints on the stability coefficients. More concretely, Model 12 assumed the first-order and higher order stability paths of all three constructs (i.e., math self-concept, math grades, and math achievement test scores) to be of equal size across measurement waves for the three school tracks. The fit of this restrictive model remained comparable to the fit of Model 11 arguing for its tenability.

Finally, Model 13 included the invariance of covariate paths meaning that the effects of the three covariates (gender, IQ, and SES) on math self-concept, math school grades, and math achievement test scores at T1 are of similar size across groups of school tracks. The model fit remained stable supporting the appropriateness of this highly restrictive model. When considering the resulting standardized path coefficients of this final model (see Table 2), it is obvious that math self-concept and math achievement (both school grades and test scores) were best predicted by their former levels given the substantial positive coefficients for the first-order stability paths. However, the significant higher order stability estimates imply that the stability of constructs does not only address consecutive waves but goes further. Second, the findings demonstrated reciprocal relations between math self-concept and math achievement both in terms of school grades and achievement test scores. The cross-lagged paths leading from math self-concept to math achievement and those leading from math achievement to math self-concept were positive and significant across all measurement waves in the three groups of students irrespective of whether achievement was operationalized by school grades or achievement test scores. Considering the effects of the covariates, gender was found to have a significant effect on math self-concept with boys displaying higher levels. Moreover, students with a higher IQ demonstrated higher levels of math self-concept whereas students' SES was found to be unrelated to math self-concept. Boys and girls were found to obtain similar school grades in math, whereas students of higher SES and higher IQ were found to earn higher grades. Regarding

Table 2
Standardized Paths Coefficients of Model 13

Time	High-achievement track	Middle-achievement track	Low-achievement track	High-achievement track	Middle-achievement track	Low-achievement track	High-achievement track	Middle-achievement track	Low-achievement track
				Stability					
	Math self-concept			Math grades			Math test scores		
T1–T2	.506*	.537*	.538*	.466*	.460*	.473*	.525*	.499*	.492*
T1–T3	.140*	.147*	.150*	.129*	.132*	.132*	.201*	.201*	.196*
T1–T4	.067*	.071*	.069*	.083*	.082*	.086*	.105*	.103*	.097*
T1–T5	.035	.036	.037	.022	.021	.022	.044*	.045	.041*
T2–T3	.522*	.518*	.525*	.416*	.430*	.417*	.445*	.468*	.462*
T2–T4	.143*	.143*	.139*	.122*	.123*	.124*	.182*	.187*	.179*
T2–T5	.068*	.067*	.068*	.080*	.078*	.080*	.089*	.095*	.088*
T3–T4	.518*	.522*	.500*	.439*	.428*	.447*	.475*	.466*	.451*
T3–T5	.143*	.141*	.140*	.131*	.124*	.131*	.182*	.184*	.173*
T4–T5	.520*	.509*	.530*	.448*	.434*	.441*	.446*	.460*	.446*
				Cross-lagged paths					
	Math grades → math self-concept			Math self-concept → math grades			Math self-concept → math test scores		
T1–T2	.088*	.090*	.093*	.047*	.049*	.049*	.056*	.054*	.056*
T2–T3	.083*	.085*	.086*	.046*	.046*	.045*	.056*	.054*	.056*
T3–T4	.087*	.087*	.087*	.046*	.046*	.046*	.059*	.056*	.055*
T4–T5	.086*	.087*	.087*	.048*	.045*	.048*	.058*	.057*	.056*
	Math test-scores → math self-concept			Math test scores → math grades			Math grades → math test scores		
T1–T2	.089*	.093*	.088*	.164*	.166*	.157*	.132*	.123*	.132*
T2–T3	.077*	.084*	.082*	.134*	.148*	.137*	.122*	.120*	.124*
T3–T4	.079*	.082*	.077*	.137*	.140*	.138*	.134*	.126*	.129*
T4–T5	.075*	.078*	.079*	.134*	.135*	.139*	.131*	.131*	.125*
				Covariates					
	Effects on math self-concept (T1)			Effects on math grades (T1)			Effects on math test scores (T1)		
Gender	.617*	.599*	.580*	.057	.058	.054	.367*	.362*	.370*
IQ	.207*	.190*	.200*	.350*	.335*	.341*	.421*	.392*	.435*
SES	.016	.017	.016	.087*	.091*	.083*	.068*	.070*	.069*

Note. Gender is coded 0 = female, 1 = male. Coefficients based on StdYX standardization within *Mplus* (i.e., standardization of independent and dependent variables) are provided for all effects except effects involving gender. For effects involving gender, coefficients based on StdY standardization are provided. SES = socioeconomic status.
* $p < .05$.

math achievement test scores, boys, students with higher IQ levels, and students of higher SES were found to demonstrate higher test scores. All these results were invariant across students from different school tracks indicating a high level of generalizability.[2]

Discussion

Even though the REM for self-concept–achievement relations has been extensively studied (Huang, 2011; Marsh & Craven, 2006), our study extends previous research and provides some of the strongest evidence for the REM so far. In essence, the present study revealed reciprocal relations between math self-concept and math achievement that were robust and generalizable in various ways.

The robustness and generalizability of the REM first becomes evident in terms of generalizability across time. The results supported the assumption of developmental equilibrium since both types of cross-lagged paths were found to be of similar sizes across the extensive time span of this study including five waves. Thus, within skill development effects and within self-enhancement effects, the effects seem to be invariant at least throughout the years of German students' mandatory secondary schooling. Interventions should therefore pursue a dual approach targeting the enhancement of both students' self-concept and achievement (Craven, Marsh, & Burnett, 2003; O'Mara, Marsh, Craven, & Debus, 2006).

The generalizability of reciprocal effects between math self-concept and math achievement, including developmental equilibrium, was further supported by considering students from different achievement tracks of the German secondary school system. Previous studies demonstrated the generalizability of the REM across different student characteristics such as age (Guay et al., 2003),

[2] To check the robustness of the findings, the same series of analyses was conducted using sampling weights. The results are reported in the online supplemental materials (see Tables S8 to S15). The results are the same as those presented herein when not using sampling weights, documenting the robustness of the analysis.

gender (Marsh, Trautwein, et al., 2005), and culture (Chen, Yeh, Hwang, & Lin, 2013; Marsh et al., 2002). This study broadens this line of research by demonstrating generalizability of the REM across students attending different achievement tracks who might experience different learning environments (Klieme & Rakoczy, 2003; Kunter et al., 2005). Practically, this finding implies that the above mentioned dual intervention approach for enhancing students' self-concept and achievement is beneficial for a wide range of students.

The robustness of reciprocal effects between math self-concept and math achievement is further reflected by the fact that the resulting pattern of relations persists when including students' gender, SES, and IQ as covariates. Hence, the assumptions of the REM even remain in place when controlling for other major determinants of students' math self-concept and math achievement.

Finally, the generalizability and robustness of the REM as demonstrated in our study addresses achievement indicators. Given that school grades and standardized achievement test scores each have their advantages and disadvantages, research benefits from including both achievement indicators in empirical studies (Marsh et al., 2014). Previous studies have indicated that the REM holds when considering school grades and achievement test scores separately (e.g., Möller et al., 2014), but only one study so far has integrated school grades and achievement test scores in a combined model (Marsh, Trautwein, et al., 2005). Given that the latter study only included two measurement waves, the present study spanning five waves is a considerable enrichment. In fact, it supported reciprocal self-concept—achievement relations for both school grades and achievement test scores in math in combined models across five measurement waves, additionally demonstrating developmental equilibrium and invariance across school tracks.

Besides providing evidence of the strong robustness and generalizability of the REM, the present study contributes to methodological approaches to the REM. It illustrates the advantage of starting with a full-forward model in which all possible paths are estimated and which thus includes first-order and higher order stability and cross-lagged paths. As exemplified in this study, such a complex model can serve as the starting point for deriving and empirically testing more parsimonious and less complex models. Accordingly, we could demonstrate that there was no additional benefit of including both first-order and higher order cross-lagged paths, but the incorporation of both first-order and higher order stability paths contributed to significantly better models. Substantively, this leads to the conclusion that self-concept and achievement are of high stability that lasts longer than across two immediately adjacent measurement waves (Marsh & O'Mara, 2008a).

A further methodological advice that can be derived from this study targets the need to include covariates which also relate to students' self-concept and achievement. Consistent with previous studies (Fredricks & Eccles, 2002; Jacobs et al., 2002; Watt, 2004), boys were found to display higher levels of math self-concept. Boys were also found to perform better on the math achievement test, but boys and girls obtained similar school grades in math. This finding corresponds to previous studies indicating that gender differences in math achievement might vary contingent upon the achievement indicator used, and that despite boys' consistent superior levels of math self-concept, gender differences in math achievement are less consistent (Hyde et al., 1990; Leahey & Guo, 2001; Lindberg, Hyde, Petersen, & Linn, 2010). Future research should also consider the situation and environmental circumstances in which students' math achievement is assessed. For example, according to the stereotype threat paradigm (Nguyen & Ryan, 2008; Steele, 1997), females' math achievement might be lower when the stereotype that girls are poorer in math than boys is activated, as compared to test situations when this gender stereotype is not prevalent.

This study followed a traditional cross-lagged modeling approach to investigate reciprocal effects between self-concept and achievement which delivers easily interpretable results and facilitates comparability across numerous previous studies on the REM that also utilized this approach (e.g., Guay et al., 2003; Marsh et al., 2007; Marsh, Trautwein, et al., 2005; Möller et al., 2011, 2014; Niepel et al., 2014; Seaton et al., 2014). However, the standard cross-lagged panel modeling approach has recently been criticized (Hamaker, Kuiper, & Grasman, 2015) mainly because of the lacking separation between the within-person level and the between-person level which would enable consideration of trait-like (i.e., stable) individual differences. Hence, it might be worthwhile to consider the application of proposed alternative models to the REM in the future. Alternative models should also be taken into account for the math achievement test. In this study, achievement test scores were scaled based on a one-parameter logistic IRT model, but alternative estimations including two-parameter models could be used in order to test the generalizability of findings. Indeed, there has been a long debate on the advantages of one-parameter relative to two-parameter IRT models (Bergan, 2013), and this debate might benefit from the application and comparison of both approaches to the same study and research question.

In light of the consistently demonstrated separation between math and verbal self-concepts (Möller, Pohlmann, Köller, & Marsh, 2009), further investigations are needed to generalize the present findings to the verbal domain. In this context, it might not only be worthwhile to study the math and verbal domains separately but to also investigate different domains simultaneously (Marsh et al., 2014). Furthermore, since only self-concept operationalized as students' perceptions of competence was considered, further variables for students' self-perceptions should be addressed such as affect self-perceptions (Arens et al., 2011; Marsh et al., 2013). Finally, beyond achievement, it might be worthwhile to take a broader range of outcome variables such as goal orientations (Seaton et al., 2014), effort (Trautwein, Lüdtke, Schnyder, & Niggli, 2006) or emotions (Pekrun, 2006) into account.

Given that this study investigated students attending the secondary school years, further long-term studies should focus on preschool or elementary school years as the present findings cannot be generalized to younger students. It can be assumed that secondary school students have established a self-concept that is sufficiently stable to impact on later achievement (Wigfield & Karpathian, 1991). This might, however, not yet be the case in preschool and elementary school years when self-enhancement effects might predominate (Arens et al., 2016; Chapman & Tunmer, 1997; Chen et al., 2013; Helmke & van Aken, 1995). Studies covering a wide time frame would be worthwhile to gain insight into the onset of reciprocal relations and developmental equilibrium in these relations.

In sum, the present study provides relevant insights into research on reciprocal relations between self-concept and achievement. In essence, the assumption of developmental equilibrium could be supported, substantiating the robustness of relations from self-concept to achievement and from achievement to self-concept across a long time interval of five waves. The robustness of reciprocal effects was further substantiated by the generalizability of the findings across achievement indicators and school tracks even when controlling for important covariates. As such, although the REM has been well established and become an inherent characteristic of the self-concept construct (Marsh & Craven, 2006), the present study has pointed out remaining important questions on reciprocal self-concept—achievement relations, delivered answers to these questions, and once again underscored the generalizability and robustness of the REM at least for secondary school students.

References

Arens, A. K., Marsh, H. W., Craven, R. G., Yeung, A. S., Randhawa, E., & Hasselhorn, M. (2016). Math self-concept in preschool children: Structure, achievement relations, and generalizability across gender. *Early Childhood Research Quarterly, 36,* 391–403. http://dx.doi.org/10.1016/j.ecresq.2015.12.024

Arens, A. K., Yeung, A. S., Craven, R. G., & Hasselhorn, M. (2011). The twofold multidimensionality of academic self-concept: Domain specificity and separation between competence and affect components. *Journal of Educational Psychology, 103,* 970–981. http://dx.doi.org/10.1037/a0025047

Baumert, J., Watermann, R., & Schümer, G. (2003). Disparitäten der Bildungsbeteiligung und des Kompetenzerwerbs: Ein institutionelles und individuelles Mediationsmodell [Disparity of educational participation and acquisition of skills: An institutional and individual model of mediation]. *Zeitschrift für Erziehungswissenschaft, 6,* 46–71. http://dx.doi.org/10.1007/s11618-003-0004-7

Becker, M., Lüdtke, O., Trautwein, U., & Baumert, J. (2006). Leistungszuwachs in Mathematik: Evidenz für einen Schereneffekt im mehrgliedrigen Schulsystem? [Achievement gains in mathematics: Evidence for differential trajectories in a tracked school system]. *Zeitschrift fur Padagogische Psychologie, 20,* 233–242. http://dx.doi.org/10.1024/1010-0652.20.4.233

Bergan, J. R. (2013). *Rasch versus Birnbaum: New arguments in an old debate*. Tucson, AZ: Assessment Technology.

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology, 53,* 371–399. http://dx.doi.org/10.1146/annurev.psych.53.100901.135233

Brookhart, S. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement, 30,* 123–142. http://dx.doi.org/10.1111/j.1745-3984.1993.tb01070.x

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: SAGE.

Brunner, M., Krauss, S., & Kunter, M. (2008). Gender differences in mathematics: Does the story need to be rewritten? *Intelligence, 36,* 403–421. http://dx.doi.org/10.1016/j.intell.2007.11.002

Calsyn, R. J., & Kenny, D. A. (1977). Self-concept of ability and perceived evaluation of others: Cause or effect of academic achievement? *Journal of Educational Psychology, 69,* 136–145. http://dx.doi.org/10.1037/0022-0663.69.2.136

Chapman, J. W., & Tunmer, W. E. (1997). A longitudinal study of beginning reading achievement and reading self-concept. *The British Journal of Educational Psychology, 67,* 279–291. http://dx.doi.org/10.1111/j.2044-8279.1997.tb01244.x

Chen, F. F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural Equation Modeling, 14,* 464–504. http://dx.doi.org/10.1080/10705510701301834

Chen, S.-K., Yeh, Y.-C., Hwang, F.-M., & Lin, S. S. J. (2013). The relationship between academic self-concept and achievement: A multicohort–multioccasion study. *Learning and Individual Differences, 23,* 172–178. http://dx.doi.org/10.1016/j.lindif.2012.07.021

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255. http://dx.doi.org/10.1207/S15328007SEM0902_5

Chmielewski, A. K., Dumont, H., & Trautwein, U. (2013). Tracking effects depend on tracking type: An international comparison of students' mathematics self-concept. *American Educational Research Journal, 50,* 925–957. http://dx.doi.org/10.3102/0002831213489843

Craven, R. G., Marsh, H. W., & Burnett, P. (2003). Cracking the self-concept enhancement conundrum. A call and blueprint for the next generation of self-concept enhancement research. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *International advances in self research: Speaking to the future* (pp. 91–126). Greenwich, CT: Information Age.

Curran, P. J., & Bollen, K. A. (2001). The best of both worlds: Combining autoregressive and latent curve models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 105–136). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10409-004

De Fraine, B., Van Damme, J., & Onghena, P. (2007). A longitudinal analysis of gender differences in academic self–concept and language achievement: A multivariate multilevel latent growth approach. *Contemporary Educational Psychology, 32,* 132–150. http://dx.doi.org/10.1016/j.cedpsych.2006.10.005

Diamantopoulos, A., & Siguaw, J. A. (2000). *Introducing LISREL*. London, UK: SAGE. http://dx.doi.org/10.4135/9781849209359

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Erikson, R., Goldthorpe, J. H., & Portocarero, L. (1979). Intergenerational class mobility in three Western European societies. *The British Journal of Sociology, 30,* 415–441. http://dx.doi.org/10.2307/589632

Fredricks, J. A., & Eccles, J. S. (2002). Children's competence and value beliefs from childhood through adolescence: Growth trajectories in two male-sex-typed domains. *Developmental Psychology, 38,* 519–533. http://dx.doi.org/10.1037/0012-1649.38.4.519

Fredricks, J. A., & Eccles, J. S. (2002). Children's competence and value beliefs from childhood through adolescence: Growth trajectories in two male-sex-typed domains. *Developmental Psychology, 38,* 519–533. http://dx.doi.org/10.1037/0012-1649.38.4.519

Frenzel, A. C., Goetz, T., Lüdtke, O., Pekrun, R., & Sutton, R. (2009). Emotional transmission in the classroom: Exploring the relationship between teacher and student enjoyment. *Journal of Educational Psychology, 101,* 705–716. http://dx.doi.org/10.1037/a0014695

Frenzel, A. C., Pekrun, R., Dicke, A. L., & Goetz, T. (2012). Beyond quantitative decline: Conceptual shifts in adolescents' development of interest in mathematics. *Developmental Psychology, 48,* 1069–1082. http://dx.doi.org/10.1037/a0026895

Frey, M. C., & Detterman, D. K. (2004). Scholastic Assessment or g? The relationship between the Scholastic Assessment Test and general cognitive ability. *Psychological Science, 15,* 373–378. http://dx.doi.org/10.1111/j.0956-7976.2004.00687.x

Furnham, A., & Monsen, J. (2009). Personality traits and intelligence predict academic school grades. *Learning and Individual Differences, 19,* 28–33. http://dx.doi.org/10.1016/j.lindif.2008.02.001

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576. http://dx.doi.org/10.1146/annurev.psych.58.110405.085530

Guay, F., Marsh, H. W., & Boivin, M. (2003). Academic self-concept and academic achievement: A developmental perspective on their causal ordering. *Journal of Educational Psychology, 95,* 124–136. http://dx.doi.org/10.1037/0022-0663.95.1.124

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods, 20,* 102–116. http://dx.doi.org/10.1037/a0038889

Hanushek, E., & Wößmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *The Economic Journal, 116,* 63–76. http://dx.doi.org/10.1111/j.1468-0297.2006.01076.x

Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision.* Göttingen, the Netherlands: Beltz Test GmbH.

Helmke, A., & van Aken, M. A. G. (1995). The causal ordering of academic achievement and self-concept of ability during elementary school: A longitudinal study. *Journal of Educational Psychology, 87,* 624–637. http://dx.doi.org/10.1037/0022-0663.87.4.624

Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica, 64,* 157–170. http://dx.doi.org/10.1111/j.1467-9574.2009.00445.x

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Huang, C. (2011). Self-concept and academic achievement: A meta-analysis of longitudinal relations. *Journal of School Psychology, 49,* 505–528. http://dx.doi.org/10.1016/j.jsp.2011.07.001

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107,* 139–155. http://dx.doi.org/10.1037/0033-2909.107.2.139

Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in children's self-competence and values: Gender and domain differences across grades one through twelve. *Child Development, 73,* 509–527. http://dx.doi.org/10.1111/1467-8624.00421

Jelicić, H., Phelps, E., & Lerner, R. M. (2009). Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology, 45,* 1195–1199. http://dx.doi.org/10.1037/a0015665

Kenny, D. A. (1975). Cross-lagged panel correlation: A test for spuriousness. *Psychological Bulletin, 82,* 887–903. http://dx.doi.org/10.1037/0033-2909.82.6.887

Klieme, E., & Rakoczy, K. (2003). Unterrichtsqualität aus Schülerperspektive: Kulturspezifische Profile, regionale Unterschiede und Zusammenhänge mit Effekten von Unterricht. In J. Baumert, C. Artelt, E. Klieme, J. Neubrand, M. Prenzel, U. Schiefele, & K.-J. Tillmann (Eds.), *PISA 2000. Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (pp. 334–359). Opladen, Germany: Leske + Budrich.

Köller, O., & Baumert, J. (2001). Leistungsgruppierungen in der sekundarstufe I. Ihre konsequenzen für die mathematikleistung und das mathematische selbstkonzept der begabung [Ability grouping at secondary level 1: Consequences for mathematics achievement and the self-concept of mathematical ability]. *Zeitschrift fur Padagogische Psychologie, 15,* 99–110.

Kunter, M., Brunner, M., Baumert, J., Klusmann, U., Krauss, S., Blum, W., . . . Neubrand, M. (2005). Der Mathematikunterricht der PISA-Schülerinnen und -Schüler. Schulformunterschiede in der Unterrichtsqualität [Quality of mathematics instruction across school types: Findings from PISA 2003]. *Zeitschrift für Erziehungswissenschaft, 4,* 502–520. http://dx.doi.org/10.1007/s11618-005-0156-8

Leahey, E., & Guo, G. (2001). Gender differences in mathematical trajectories. *Social Forces, 80,* 713–732. http://dx.doi.org/10.1353/sof.2001.0102

LeTendre, G. K., Hofer, B. K., & Shimizu, H. (2003). What is tracking? Cultural expectations in the United States, Germany, and Japan. *American Educational Research Journal, 40,* 43–89. http://dx.doi.org/10.3102/00028312040001043

Lietz, P. (2006). A meta-analysis of gender differences in reading achievement at the secondary school level. *Studies in Educational Evaluation, 32,* 317–344. http://dx.doi.org/10.1016/j.stueduc.2006.10.002

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin, 136,* 1123–1135. http://dx.doi.org/10.1037/a0021276

Little, R., & Rubin, D. (2002). *Statistical analysis with missing data.* New York, NY: Wiley. http://dx.doi.org/10.1002/9781119013563.

Liu, X. (2010). *Using and developing measurement instruments in science education: A Rasch modeling approach.* Charlotte, NC: Information Age Publishing.

Marsh, H. W. (1990). The causal ordering of academic self-concept and academic achievement: A multiwave, longitudinal panel analysis. *Journal of Educational Psychology, 82,* 646–656. http://dx.doi.org/10.1037/0022-0663.82.4.646

Marsh, H. W. (2007). *Self-concept theory, measurement and research into practice: The role of self-concept in educational psychology.* Leicester, UK: British Psychological Society.

Marsh, H. W., Abduljabbar, A. S., Abu-Hilal, M., Morin, A. J. S., Abdelfattah, F., Leung, K. C., . . . Parker, P. (2013). Factor structure, discriminant and convergent validity of TIMSS math and science motivation measures: A comparison of USA and Saudi Arabia. *Journal of Educational Psychology, 105,* 108–128. http://dx.doi.org/10.1037/a0029907

Marsh, H. W., Byrne, B. M., & Yeung, A. S. (1999). Causal ordering of academic self-concept and achievement: Reanalysis of a pioneering study and revised recommendations. *Educational Psychologist, 34,* 155–167. http://dx.doi.org/10.1207/s15326985ep3403_2

Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective. Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science, 1,* 133–163. http://dx.doi.org/10.1111/j.1745-6916.2006.00010.x

Marsh, H. W., Craven, R. G., Parada, R. H., Guo, J., Dicke, T., & Abduljabbar, A. S. (2016). Temporal ordering effects of adolescent depression, relational aggression and victimization over six waves: Fully latent teciprocal rffects models. *Developmental Psychology.* Advance online publication. http://dx.doi.org/10.1037/dev0000241

Marsh, H. W., Dowson, M., Pietsch, J., & Walker, R. (2004). Why multicollinearity matters: A reexamination of relations between self-efficacy, self-concept, and achievement. *Journal of Educational Psychology, 96,* 518–522. http://dx.doi.org/10.1037/0022-0663.96.3.518

Marsh, H. W., Gerlach, E., Trautwein, U., Lüdtke, O., & Brettschneider, W.-D. (2007). Longitudinal study of preadolescent sport self-concept and performance: Reciprocal effects and causal ordering. *Child Development, 78,* 1640–1656. http://dx.doi.org/10.1111/j.1467-8624.2007.01094.x

Marsh, H. W., & Hau, K.-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *Journal of Experimental Education, 64,* 364–390. http://dx.doi.org/10.1080/00220973.1996.10806604

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit evaluation in structural equation modeling. In A. Maydeu-Olivares & J. McArdle (Eds.), *Contemporary psychometrics. A Festschrift for Roderick P. McDonald.* Mahwah, NJ: Lawrence Erlbaum.

Marsh, H. W., Hau, K.-T., & Kong, K. W. (2002). Multilevel causal ordering or academic self-concept and achievement: Influence of language of instruction (English vs. Chinese) for Hong Kong students. *American Educational Research Journal, 39,* 727–763. http://dx.doi.org/10.3102/00028312039003727

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to cutoff values for fit indexes and dangers in overgeneralizing Hu & Bentler's (1999). *Structural Equation Modeling, 11,* 320–341. http://dx.doi.org/10.1207/s15328007sem1103_2

Marsh, H. W., Kuyper, H., Seaton, M., Parker, P. D., Morin, A. J. S., Möller, J., & Abduljabbar, A. S. (2014). Dimensional comparison theory: An extension of the internal/external frame of reference effect on academic self-concept formation. *Contemporary Educational Psychology, 39,* 326–341. http://dx.doi.org/10.1016/j.cedpsych.2014.08.003

Marsh, H. W., & O'Mara, A. (2008a). Reciprocal effects between academic self-concept, self-esteem, achievement, and attainment over seven adolescent years: Unidimensional and multidimensional perspectives of self-concept. *Personality and Social Psychology Bulletin, 34,* 542–552. http://dx.doi.org/10.1177/0146167207312313

Marsh, H. W., & O'Mara, A. J. (2008b). Self-concept is as multidisciplinary as it is multidimensional. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *Self-processes, learning, and enabling human potential. Dynamic new approaches* (pp. 87–115). Charlotte, NC: Information Age.

Marsh, H. W., Pekrun, R., Lichtenfeld, S., Guo, J., Arens, A. K., & Murayama, K. (2016). Breaking the double-edged sword of effort/trying hard: Developmental equilibrium and longitudinal relations among effort, achievement, and academic self-concept. *Developmental Psychology, 52,* 1273–1290. http://dx.doi.org/10.1037/dev0000146

Marsh, H. W., Pekrun, R., Parker, P. D., Murayama, K., Guo, J., Dicke, T., & Lichtenfeld, S. (2016). Long-term positive effects of repeating a year in school: Six-year longitudinal study of self-beliefs, anxiety, social relations, school grades, and test scores. *Journal of Educational Psychology*. Advance online publication. http://dx.doi.org/10.1037/edu0000144

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development, 76,* 397–416. http://dx.doi.org/10.1111/j.1467-8624.2005.00853.x

Marsh, H. W., & Yeung, A. S. (1998). Longitudinal structural equation models of academic self-concept and achievement: Gender differences in the development of math and English constructs. *American Educational Research Journal, 35,* 705–738. http://dx.doi.org/10.3102/00028312035004705

Marshall, S. L., Parker, P. D., Ciarrochi, J., & Heaven, P. C. L. (2014). Is self-esteem a cause or consequence of social support? A 4-year longitudinal study. *Child Development, 85,* 1275–1291. http://dx.doi.org/10.1111/cdev.12176

Matteucci, M., & Mignani, S. (2011). Gender differences in performance in mathematics at the end of lower secondary school in Italy. *Learning and Individual Differences, 21,* 543–548. http://dx.doi.org/10.1016/j.lindif.2011.03.001

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research, 95,* 203–213. http://dx.doi.org/10.1080/00220670209596593

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* New York, NY: Routledge.

Möller, J., Pohlmann, B., Köller, O., & Marsh, H. W. (2009). Meta-analytic path analysis of the internal/external frame of reference model of academic achievement and academic self-concept. *Review of Educational Research, 79,* 1129–1167. http://dx.doi.org/10.3102/0034654309337522

Möller, J., Retelsdorf, J., Köller, O., & Marsh, H. W. (2011). The reciprocal internal/external frame of reference model: An integration of models of relations between academic achievement and self-concept. *American Educational Research Journal, 48,* 1315–1346. http://dx.doi.org/10.3102/0002831211419649

Möller, J., Zimmermann, F., & Köller, O. (2014). The reciprocal internal/external frame of reference model using grades and test scores. *The British Journal of Educational Psychology, 84,* 591–611. http://dx.doi.org/10.1111/bjep.12047

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report: IEA's progress in international reading literacy study in primary schools in 40 countries.* Chestnut Hill, MA: Boston College.

Murayama, K., Pekrun, R., Lichtenfeld, S., & Vom Hofe, R. (2013). Predicting long-term growth in students' mathematics achievement: The unique contributions of motivation and cognitive strategies. *Child Development, 84,* 1475–1490. http://dx.doi.org/10.1111/cdev.12036

Murayama, K., Pekrun, R., Suzuki, M., Marsh, H. W., & Lichtenfeld, S. (2016). Don't aim too high for your kids: Parental over-aspiration undermines students' learning in mathematics. *Journal of Personality and Social Psychology, 111,* 766–779. http://dx.doi.org/10.1037/pspp0000079

Muthén, B., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 267–316). Washington, DC: American Sociological Association. http://dx.doi.org/10.2307/271070

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Nguyen, H. H., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology, 93,* 1314–1334. http://dx.doi.org/10.1037/a0012702

Niepel, C., Brunner, M., & Preckel, F. (2014). The longitudinal interplay of students' academic self-concepts and achievements within and across domains: Replicating and extending the reciprocal internal/external frame of reference model. *Journal of Educational Psychology, 106,* 1170–1191. http://dx.doi.org/10.1037/a0036307

Nowell, A., & Hedges, L. V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance, and extreme scores. *Sex Roles, 39,* 21–43. http://dx.doi.org/10.1023/A:1018873615316

O'Mara, A. J., Marsh, H. W., Craven, R. G., & Debus, R. L. (2006). Do self-concept interventions make a difference? A synergistic blend of construct validation and meta-analysis. *Educational Psychologist, 41,* 181–206. http://dx.doi.org/10.1207/s15326985ep4103_4

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review, 18,* 315–341. http://dx.doi.org/10.1007/s10648-006-9029-9

Pekrun, R., Lichtenfeld, S., Marsh, H. W., Murayama, K., & Goetz, T. (in press). Achievement emotions and academic performance: Longitudinal models of reciprocal effects. *Child Development.*

Pekrun, R., vom Hofe, R., Blum, W., Frenzel, A. C., Goetz, T., & Wartha, S. (2007). Development of mathematical competencies in adolescence: The PALMA longitudinal study. In M. Prenzel (Ed.), *Studies on the educational quality of schools* (pp. 17–37). Münster, Germany: Waxmann.

Raykov, T. (2009). Evaluation of scale reliability for unidimensional measures using latent variable modeling. *Measurement & Evaluation in Counseling & Development, 42,* 223–232. http://dx.doi.org/10.1177/0748175609344096

Retelsdorf, J., Köller, O., & Möller, J. (2014). Reading achievement and reading self-concept. Testing the reciprocal effects model. *Learning and Instruction, 29,* 21–30. http://dx.doi.org/10.1016/j.learninstruc.2013.07.004

Seaton, M., Parker, P., Marsh, H. W., Craven, R. G., & Yeung, A. S. (2014). The reciprocal relations between self-concept, motivation and achievement: Juxtaposing academic self-concept and achievement goal

orientations for mathematics success. *Educational Psychology, 34,* 49–72. http://dx.doi.org/10.1080/01443410.2013.825232

Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Self-concept: Validation of construct interpretations. *Review of Educational Research, 46,* 407–441. http://dx.doi.org/10.3102/00346543046003407

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75,* 417–453. http://dx.doi.org/10.3102/00346543075003417

Skaalvik, S., & Skaalvik, E. M. (2004). Gender differences in math and verbal self-concept, performance expectations, and motivation. *Sex Roles, 50,* 241–252. http://dx.doi.org/10.1023/B:SERS.0000015555.40976.e6

Spinath, F. M., Spinath, B., & Plomin, R. (2008). The nature and nurture of intelligence and motivation in the origins of sex differences in elementary school achievement. *European Journal of Personality, 22,* 211–229. http://dx.doi.org/10.1002/per.677

Steele, C. M. (1997). A threat in the air. How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613–629. http://dx.doi.org/10.1037/0003-066X.52.6.613

Trautwein, U., Lüdtke, O., Schnyder, I., & Niggli, A. (2006). Predicting homework effort: Support for a domain-specific, multilevel homework model. *Journal of Educational Psychology, 98,* 438–456. http://dx.doi.org/10.1037/0022-0663.98.2.438

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist, 39,* 111–131. http://dx.doi.org/10.1207/s15326985ep3902_3

Van de gaer, E., Pustjens, H., Van Damme, J., & De Munter, A. (2008). Mathematics participation and mathematics achievement across secondary school: The role of gender. *Sex Roles, 59,* 568–585.

Vom Hofe, R., Kleine, M., Blum, W., & Pekrun, R. (2005). On the role of "Grundvorstellungen" for the development of mathematical literacy. First results of the longitudinal study PALMA. *Mediterranean Journal for Research in Mathematics Education, 4,* 67–84.

Vom Hofe, R., Pekrun, R., Kleine, M., & Götz, T. (2002). Projekt zur Analyse der Leistungsentwicklung in Mathematik (PALMA): Konstruktion des Regensburger Mathematikleistungstests für 5.-10. Klassen [Project for the Analysis of Learning and Achievement in Mathematics (PALMA): Development of the Regensburg Mathematics Achievement Test for grades 5 to 10]. *Zeitschrift fur Padagogik, 45,* 83–100.

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus.* Chichester, UK: Wiley. http://dx.doi.org/10.1002/9781118356258

Watt, H. M. G. (2004). Development of adolescents' self-perceptions, values, and task perceptions according to gender and domain in 7th- through 11th-grade Australian students. *Child Development, 75,* 1556–1574. http://dx.doi.org/10.1111/j.1467-8624.2004.00757.x

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives, 4,* 10–18. http://dx.doi.org/10.1111/j.1750-8606.2009.00110.x

Wigfield, A., & Karpathian, M. (1991). Who am I and what can I do? Children's self- concepts and motivation in achievement solutions. *Educational Psychologist, 26,* 223–261. http://dx.doi.org/10.1080/00461520.1991.9653134

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. I. Hershberger (Eds.), *The new rules of measurement: What every educator and psychologist should know.* Hillsdale, NJ: Lawrence Erbaum.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACERConQuest Version 2: Generalised item response modelling software.* Camberwell, Australia: Australian Council for Educational Research.

Zimmermann, F., Schütte, K., Taskinen, P., & Köller, O. (2013). Reciprocal effects between adolescent externalizing problems and measures of achievement. *Journal of Educational Psychology, 105,* 747–761. http://dx.doi.org/10.1037/a0032793

Received July 7, 2015
Revision received September 19, 2016
Accepted September 23, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 5, 635–652

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000172

In Peer Matters, Teachers Matter: Peer Group Influences on Students' Engagement Depend on Teacher Involvement

Justin W. Vollet, Thomas A. Kindermann, and Ellen A. Skinner
Portland State University

This study focused on the joint effects of teachers and peer groups as predictors of change in students' engagement during the first year of middle school, when the importance of peer relationships normatively increases and the quality of teacher–student relationships typically declines. To explore cumulative and contextualized joint effects, the study utilized 3 sources of information about an entire cohort of 366 sixth graders in a small town: Peer groups were identified using sociocognitive mapping; students reported on teacher involvement; and teachers reported on each student's engagement. Consistent with models of cumulative effects, peer group engagement and teacher involvement each uniquely predicted changes in students' engagement. Consistent with contextualized models suggesting differential susceptibility, peer group engagement was a more pronounced predictor of changes in engagement for students who experienced relatively low involvement from teachers. These peer effects were positive or negative depending on the engagement versus disaffection of each student's peer group. Person-centered analyses also revealed cumulative and contextualized effects. Most engaged were students who experienced support from both social partners; steepest engagement declines were found when students affiliated with disaffected peers and experienced teachers as relatively uninvolved. High teacher involvement partially protected students from the motivational costs of affiliating with disaffected peers, and belonging to engaged peer groups partially buffered students' engagement from the effects of low teacher involvement. These findings suggest that, although peer groups and teachers are each important individually, a complete understanding of their contributions to students' engagement requires the examination of their joint effects.

Keywords: student engagement, peer influence, teacher influence, differential susceptibility, joint effects

The construct of academic engagement has sparked considerable enthusiasm in both research and educational communities in recent decades for three reasons. First, engagement is a robust predictor of academic success, showing links to students' learning (Blondal & Adalbjarnardottir, 2012), retention and graduation (Finn, 1989), and educational achievement and attainment (Finn & Zimmer, 2012), across all educational grade levels (Hughes & Kwok, 2007; Shernoff, Csikszentmihalyi, Shneider, & Shernoff, 2003; Skinner, Kindermann, Connell, & Wellborn, 2009; Ullah & Wilson, 2007). Second, engagement seems to offer some protection from developmentally risky behaviors, such as drop-out and delinquency (Fall & Roberts, 2012; Li & Lerner, 2011; Wang & Fredricks, 2014; Wang & Peck, 2013), especially during early and middle adolescence, when these behaviors are otherwise normatively on the rise. And third, studies indicate that engagement is malleable and so open to intervention efforts (e.g., Anderson, Christenson, Sinclair, & Lehr, 2004).

In recent years, a fourth source of enthusiasm about the construct of engagement has emerged, as motivational researchers have begun to explore the overlap between engagement, a construct grounded in educational, psychological, and sociological traditions targeting antidotes to student drop-out (Finn, 1989; Newmann, 1992; Rumberger & Rotermund, 2012), and research on achievement motivation, an area of study grounded in the older and broader field of motivation (Deci, 1992; Weiner, 1990). Leaders in the field have recently suggested that student engagement may be considered an outcome of motivation and, as a result, research on engagement is now included in definitive reviews of motivational research (e.g., Wentzel & Miele, 2016; Wigfield et al., 2015). Some motivational theorists even argue that classroom engagement, defined as students' ongoing, active, and energized participation in academic tasks, is a potential marker of a motivated state, and so can be considered an observable manifestation of the energy and persistence generated by underlying motivation (R. M. Ryan & Deci, 2009; Skinner, Kindermann, Connell, et al., 2009; Wang & Degol, 2014). Viewing engagement from a motivational perspective opens up the possibility that many of the factors already established as important predictors of motivational development may also serve to support students' classroom engagement (Reeve, 2012; Skinner, 2016; Wang & Eccles, 2012).

In fact, much of the research examining the ways in which students' engagement can be shaped by their interpersonal relationships in school has relied on motivational accounts of the influences of teachers and peers (Martin & Dowson, 2009; Wentzel, 2009a, 2009b). To date, studies have largely concentrated on the role of teachers (Quin, in press; Roeser, Eccles, & Sameroff,

This article was published Online First January 2, 2017.

Justin W. Vollet, Thomas A. Kindermann, and Ellen A. Skinner, Department of Psychology, Portland State University.

Correspondence concerning this article should be addressed to Justin W. Vollet, Department of Psychology, Portland State University, PO Box 751, Portland, OR 97207-0751. E-mail: jwv@pdx.edu

2000; Wentzel, 1997), but in recent years, research has begun to expand to include friends, classmates, and peer groups (Wentzel & Ramani, 2016). Up until now, however, few studies have looped back to examine the role of peer groups in combination with students' relationships with teachers, despite previous work that documents the centrality of teachers to student motivation and engagement (Quin, in press; Wentzel, 2009b). Guided by ecological models of schools as complex social systems (Bronfenbrenner & Morris, 2006), the purpose of the current study was to provide a more contextualized view of peer group contributions to academic engagement, by considering how their impact could be shaped by students' relationships with teachers, specifically, students' experiences of their teachers' involvement.

A Social–Ecological Model of the Impact of Teachers and Peers on Student Engagement

The present study was framed by an ecological perspective. This framework suggests that complex social ecologies, like schools, can be conceptualized as multifaceted systems that contain multiple subsystems, and these subsystems work together to shape student development. If peer groups represent one such subsystem and teacher–student relationships represent a second, then an ecological perspective suggests that it may be important to examine them jointly, and posits two primary ways in which they can work together. First, teachers and peer groups may exert "cumulative" or additive influences, in which the contributions of social partners accrue in their effects and in which, despite some overlap, each may provide essential supports that the other cannot. Second, the influences of teachers and peer groups may be "contextualized" or interactive, in that the impact of one set of social partners may depend on the nature of the other. Many kinds of contextualized interactions among subsystems can be imagined, such as *compensatory* effects, in which support from one social partner protects students from the negative impact of the other, or *amplifying* effects, in which the positive or negative attributes of one social partner magnify the corresponding positive or negative effects of the other. Such a perspective suggests that the effects of peer groups may be both cumulative and contextualized—peers not only provide unique supports to academic engagement, but their effects also depend on the quality of students' relationships with teachers. Although an ecological perspective highlights the possibility of joint effects of teachers and peers, it does not specify how and why teacher involvement might temper the influence of peer groups. For guidance on these more specific questions, we turned to Self-Determination Theory (SDT; Connell & Wellborn, 1991; Deci & Ryan, 1985) and theories of Stage–Environment Fit (SEF; Eccles et al., 1993; Eccles & Roeser, 2009).

Teacher Involvement and Student Engagement

SDT posits that all people, including students in classrooms, have fundamental psychological needs for relatedness, competence, and autonomy. When those needs are fulfilled by participation in an enterprise, like school, individuals will more constructively take part in the activities of that enterprise; for example, students will engage more fully with learning activities in classrooms and cooperate more willingly with school rules (R. M. Ryan & Deci, 2009). Consistent with this theory, decades of research have shown that students evince greater engagement when teachers provide higher levels of support for students' motivational needs, including warmth, pedagogical caring (Wentzel, 1997), closeness (Hamre & Pianta, 2001), acceptance (Wentzel, 1994), help, direction (A. M. Ryan & Shin, 2011), involvement, provision of structure, and autonomy support (Klem & Connell, 2004; Skinner & Belmont, 1993; see Quin, in press, for a review). A primary pathway through which teacher motivational support shapes engagement is by helping students feel more efficacious, autonomous, welcome, and safe, and to better internalize educational values (Connell & Wellborn, 1991; Deci & Ryan, 1985; Reeve, 2012; Skinner & Belmont, 1993; Wentzel, 1999, 2009b; Wigfield et al., 2015).

Although studies have identified a wide band of teacher behaviors that promote student motivation and engagement, research suggests that central among them is teacher provision of pedagogical caring (Wentzel, 1997) or involvement (Skinner & Belmont, 1993), which focuses on a constellation of teacher behaviors, including warmth, affection, and enjoyment, that mark a close and caring teacher–student relationship. One pathway through which teacher involvement seems to support student motivation and engagement is by fostering students' sense of belonging (Osterman, 2000; Goodenow, 1993), relatedness (Furrer & Skinner, 2003), or attachment to school (Libbey, 2004). According to SDT, relatedness to teachers (and other social partners) acts like "psychological glue" that connects students to school and promotes their engagement. From this perspective, high-quality student–teacher relationships, characterized by involvement and affection, are a foundation upon which the development of motivation and engagement depend (Eccles & Roeser, 2009; Reeve, 2012; Wentzel, 2009a; Wigfield et al., 2015).

At the same time, SEF alerts researchers to the importance of early adolescence and the transition to middle school as a time when students' needs for relatedness may become increasingly strained. Just when young adolescents are testing their fledgling independence from parents by reaching out for closer connections to peers and adults outside the home, like teachers, the quality of students' relationships with teachers begins to decline (according to reports from both students and teachers; Wigfield et al., 2015). These declines may be due at least in part to organizational changes in which students shift from having few to many teachers per day, making it more difficult to build close connections (Eccles & Roeser, 2009). SEF highlights this stage–environment mismatch, and suggests that declines in the quality of teacher–student relationships, which parallel declines in student engagement, may be a major contributor to losses in engagement and motivation over the transition to middle school (Eccles & Roeser, 2009; Wigfield et al., 2015). Because of its centrality in promoting student motivation and its well-documented decline at the middle school transition, the current study focused on the role of teacher involvement, specifically, students' experiences of their teachers as involved (affectionate, caring, and dependable) as a potential predictor of students' engagement in the classroom.

Challenges to Examining Teachers' Involvement

While studies converge on the importance of teacher involvement to student engagement and motivation, researchers who aim to assess teachers' influence during middle school still face distinct

challenges (Wentzel, 2009b). Although the identification of students' teachers may be straightforward, identifying those teachers who are best positioned to influence students' engagement is not. This is a particularly thorny issue for research in middle schools, where students interact with multiple teachers throughout the day. To overcome this problem, many researchers use measures that assess students' experiences of their teachers in general (e.g., Wang & Eccles, 2012), thereby allowing students themselves to aggregate the most salient influences. In support of this practice, researchers have used questionnaires tapping students' perceptions of teacher involvement that include the stem "My teacher . . ." in longitudinal studies from elementary school through middle school (e.g., De Laet, et al., 2015; Skinner & Belmont, 1993). Evidence of the functioning of these scores over time indicates that, at least under these conditions, such measures maintain their key psychometric and validity characteristics (Skinner, Kindermann, & Furrer, 2009; Skinner, Zimmer-Gembeck, Connell, Eccles, & Wellborn, 1998).

Peer Groups and Student Engagement

Multiple strands of research have converged on the conclusion that classmates and friends also play a significant role in student motivation and engagement in school (Wentzel, 2009a). Although much of this research has focused on close, reciprocated friendships as sources of enjoyment and correlates of success in school (Altermatt & Pomerantz, 2005; Berndt, Hawkins, & Jiao, 1999; Hallinan & Williams, 1990; Ladd, 1990), a growing number of studies have examined the role of naturally occurring peer groups. This work explores the proposition that one way peers influence student engagement, motivation, and achievement is through proximal processes that occur in frequent social interactions within self-selected groups of peers (Kindermann, 2007; A. M. Ryan, 2000, 2001). A key idea is that participation in groups of peers who are engaged or disaffected from school has the potential, in addition to the contributions of friendship relationships and dyadic interactions with peers, to impact students' own emotional and behavioral engagement in the classroom (Kindermann & Skinner, 2012). Theories of peer group influence have suggested that their effects may be conveyed through multiple channels. They may be transmitted directly, through mechanisms of socialization, including modeling, reinforcement, encouragement, or pressure to conform to group norms (Altermatt & Pomerantz, 2005; Harris, 1995; Kindermann, 2003; Lynch, Lerner, & Leventhal, 2013), as well as indirectly, for example, by fulfilling needs for relatedness (Anderman & Anderman, 1999; Nelson & DeBacker, 2008; Furrer & Skinner, 2003) or providing academic help and support (Lempers & Clark-Lempers, 1992; Wentzel & Watkins, 2011).

Peer groups, which can be viewed as largely self-selected social contexts, provide opportunities for dyadic interactions and the formation of friendship relationships with similar peers (Kindermann & Skinner, 2012). Because peer groups tend to be selected based on similarity (i.e., homophily), such groups can create a more concentrated or intensified local context that, in the case of engagement and disaffection, may surround students who are already engaged with a higher concentration of engaged peers, and expose students who are already somewhat disaffected to a higher concentration of disaffected peers, thus potentially amplifying individuals' initial motivational states over time. For example, studies have shown that peer groups' average engagement levels at the beginning of the school year are small but robust predictors of changes in students' teacher-reported engagement over the year, during both elementary and middle school (Kindermann, 1993, 2007). In the same vein, A. M. Ryan (2001) found that middle school students who affiliated with peers who disliked school showed the steepest declines in their own enjoyment of school. Because of their potential importance to the development of students' engagement, the current study focused on the role of peer groups, specifically. the extent to which the members of an individual student's peer group were engaged versus disaffected with academic activities in the classroom.

Challenges to the Study Peer Groups

While a growing number of studies have pointed to the important role peer groups play in the development of student engagement and motivation, they have also highlighted two key challenges to investigating their effects. First, it can be difficult to reliably identify children's peer groups in naturalistic contexts, like schools. Natural peer groups consist of the agemates with whom children regularly interact. Such groups are hard to define because they are self-organized, evolve rapidly, and are often overlapping. To address this challenge, the current study used sociocognitive mapping (SCM; Cairns, Perrin, & Cairns, 1985), which uses students themselves as expert observers of group interactions. Because students have the opportunity to witness schoolmates' public exchanges every day, such "insider" observations afford the most complete access to information about naturally occurring peer groups. Another advantage of relying on multiple observers is that it allows for an assessment of the level of agreement between reporters. Furthermore, unlike self-reports, which require near complete participation (otherwise each nonparticipating child is also missing as a potential peer group member of participating children), the accuracy of SCM is less affected by participation rates, because other reporters typically include missing group members. In fact, Cairns and Cairns (1994) estimated a criterion such that, when the sample of reporters is relatively representative, reports from slightly more than half the student body are sufficient to yield reliable networks.

Once the members of each student's peer groups have been identified, a second challenge is to figure out how to capture meaningful characteristics of groups. One method, used in the present study, is to create peer profile scores for each child, by identifying the members of a target child's peer group, and then combining measures of key characteristics obtained for each member (Kindermann, 1993, 1996; Kurdek & Sinclair, 2000; A. M. Ryan, 2001). Peer profiles of engagement can be calculated for a given student by averaging the engagement scores of each member of his or her peer group. In the current study, SCM was used to identify the members of each child's peer groups, and peer profiles of engagement versus disaffection were used to capture the motivational composition of each child's local peer context, with the expectation that these profiles might predict changes in individual student's engagement over the school year.

Studies of Joint Effects of Teachers and Peers

As research on peer group influences has begun to accumulate, findings seem to converge on their potential importance to student

motivation and engagement (Wentzel & Muenks, 2016). However, few of these studies have tried to incorporate the impact of the other major social partner in the classroom, namely, teachers. To date, only seven studies have examined the joint effects of teachers and peers on student academic engagement, motivation, or success. To guide our own examination of the interplay between teachers and peers, we built on the few studies that have begun to incorporate the effects of both social partners, looking carefully at the attributes they targeted and how they analyzed different configurations of these relationships.

Evidence for Cumulative Effects

Of the seven studies of joint effects, four found evidence for only cumulative effects, in which peers contributed uniquely to student engagement over and above the effect of teachers. In a large sample of 13-year-old students in Norway, Danielsen, Wiium, Wilhelmsen, and Wold (2010) found that perceptions of support from teachers (i.e., friendliness and fairness) and peers (i.e., classmates' acceptance, kindness and helpfulness, and sense of togetherness) each uniquely predicted students' self-reported academic initiative (tapped using items such as "I challenge myself when I am doing schoolwork") at the individual level; interactions were not examined. In a second study, Wentzel, Battle, Russell, and Looney (2010) analyzed the extent to which middle schoolers' perceptions of four kinds of supports from teachers and peers (expectations for academic engagement and positive social behavior, provisions of help, safety, and emotional nurturing) were related to school motivation. Multiple regressions, controlling for sex, grade level, and teacher, revealed that all four of the teacher supports uniquely predicted student self-reported academic motivation. When peer supports were entered in the last step, both peer expectations and help were also unique predictors, although none of the interactions between corresponding teacher and peer supports were significant.

In a third study, De Laet and colleagues (2015) investigated whether relationship qualities of teachers (including global support and conflict) and peers (including popularity and acceptance) jointly predicted the development of children's behavioral engagement from Grades 4 to 6. An additive model showed the best fit to the data, indicating that high and increasing levels of teacher support and high levels of peer acceptance (but not teacher conflict or peer popularity) contributed independently to counteract the normative declines in children's behavioral engagement. Analyses of moderation and mediation were conducted but none were found. In a fourth study, Wang and Eccles (2012) examined growth curves of behavioral, emotional, and cognitive engagement from Grades 7 to 11, which they assessed using student reports of school compliance, identification with school, and subjective value of learning, respectively. Although, in general, all three dimensions of engagement showed the normative declines typical for these ages/grades, students' reports of peer support predicted more favorable trajectories, that is, less-steep declines, over and above the effects of support from teachers (and parents). No two-way interactions with support from teachers (or parents) qualified the protective contributions of peer support, but these effects were more pronounced for the trajectories of emotional and cognitive engagement of African American students. Moreover, protective effects of peer support on behavioral engagement were found only for students who reported having more prosocial friends. For students reporting more antisocial friends, higher levels of peer support actually exacerbated declines in engagement.

Evidence for Contextualized Effects

Three additional studies found evidence for both cumulative and contextualized effects. All three used either pattern-oriented or person-centered analyses to examine groups of students who differed in their profiles of relationships with peers and teachers. In an early study, Furrer and Skinner (2003) examined third- through sixth-grade students' reports of their relatedness to teachers and peers (as well as parents) as predictors of self- and teacher-reported behavioral and emotional engagement. Using median splits, they created groups of students with high relatedness to none, one, two, or all three of these social partners. Group comparisons revealed that high relatedness to teachers could compensate for low relatedness to peers, but high relatedness to peers could not compensate for low relatedness to teachers. In a second study, Davidson and colleagues (2010) focused on school adjustment as the target outcome (assessed as an aggregate of academic skills and self-concept, school bonding, loneliness, and self-worth), and used latent profile analysis to identify students with different patterns of relatedness to teachers and peers (based on teacher-reported teacher–student closeness, peer-nominated peer social preference, and self-reports of perceived peer competence). Three groups were distinguished: (a) high relatedness (high on all three indicators); (b) low relatedness (low on all three indicators); and (c) peer oriented (medium teacher–student closeness combined with high peer social preference and perceived peer competence). Although adjustment was lowest for students who reported low relatedness to both peers and teachers at the beginning of the sixth grade, students from both high-relatedness and peer-oriented groups evinced positive academic adjustment, suggesting that positive relationships with peers might be sufficient to support adjustment, even without highly supportive relationships with teachers. Finally, Raufelder, Jagenow, Drury, and Hoferichter (2013) used latent class analysis to identify four groups of students, namely, those who reported that their academic motivation was more dependent on teachers, more dependent on peers, dependent on both, or dependent on neither. Interestingly, the largest group of students consisted of those who saw their academic motivation as primarily dependent on peers, followed by students who saw their motivation as simultaneously peer and teacher dependent. Degree of membership in all four clusters was associated with several markers of motivation, including academic drive, learning goals, striving for academic success, and avoidance of academic failure.

Critique of Current Studies of Joint Effects of Peers and Teachers

Taken together, these studies provide evidence for two complementary perspectives on how teachers and peers work together to shape students' academic engagement, motivation, and adjustment. On the one hand, all seven found evidence of cumulative effects of peers and teachers, whether studies used variable-centered analyses (Danielsen et al., 2010; De Laet et al., 2015; Wang & Eccles, 2012; Wentzel et al., 2010) or person-centered analyses (Davidson et al., 2010; Furrer & Skinner, 2003; Raufelder

et al., 2013), suggesting that peers play an important role in their own right, a role not completely filled by teachers, no matter how much support they provide. However, studies did not converge on whether joint effects are also contextualized, that is, whether peer group effects are qualified to some extent by students' relationships with their teachers. In fact, three studies explicitly tested for interactions, but did not find them (De Laet et al., 2015; Wang & Eccles, 2012; Wentzel et al., 2010). In trying to explain these differences, it may be significant that in two of these three studies (Wang & Eccles, 2012; Wentzel et al., 2010), researchers relied on student-report measures to tap all three of the key constructs, namely, teacher support, peer support, and motivational outcomes. It is possible that common-method variance makes it more difficult to disentangle the differential effects of the three parties involved. In the current study, information about each player was provided by separate sources. Such separation may facilitate the detection of these more complex interactive effects. Consistent with this notion, other studies that used multiple independent reporters also uncovered interactive effects (e.g., Davidson et al., 2010).

A second factor contributing to differing patterns of results in previous studies could be the specific characteristics of teachers and peers that researchers targeted for investigation. Studies were relatively consistent in their selection of teacher factors. All seven studies focused on the social–emotional qualities of student–teacher relationships that have been shown to predict student motivation and engagement (such as involvement, closeness, friendliness, fairness, positive expectations, and provision of help, safety, and emotional nurturing; Sabol & Pianta, 2012; Wentzel, 2009b). However, studies varied widely in the peer attributes they targeted. Some included student ratings of general support from peers and teachers (Wang & Eccles, 2012) or feelings of relatedness to both partners (Furrer & Skinner, 2003). Some examined specific qualities of peer relationships that were strictly parallel to those examined in teachers (e.g., Wentzel et al., 2010). Other studies selected peer characteristics that were not exactly the same as those of teachers, but were also in the general domain of social–emotional relationship qualities (i.e., classmates' acceptance, kindness, helpfulness, and togetherness; Danielsen et al., 2010). Finally, some researchers focused on key markers of overall positive functioning in the peer domain, such as peer-nominated popularity, acceptance, or social preference (De Laet et al., 2015; Davidson et al., 2010).

In the current study, consistent with other researchers of joint effects, we examined the social emotional quality of students' relationships with teachers as predictors of their engagement. However, we differed from all previous studies in the peer characteristics we decided to target. Instead of examining peer relatedness or peer support, which have been the focus of previous studies, we targeted the engagement profiles of students' naturally occurring peer groups. We reasoned that, unlike teachers, peers do not typically have the goal of promoting a student's motivation, so their efficacy may not reside in the quality of their relationships or the support they provide. Instead, peers may shape engagement through the power of joint activity, that is, students' own engagement may be buoyed by participating as an active member of a group of enthusiastically engaged agemates who enjoy and work hard at learning activities. In contrast, trying to complete learning activities within local contexts of disaffected peers who may be passive, bored, frustrated, or discouraged can exert a downward pressure on students' own engagement, and so eventually reinforce or intensify their own disaffection.

The Interplay of Teacher and Peer Group Influences in School

The current study attempted to build on previous studies of joint effects, integrating them using an ecological framework focused on motivational theories, and strengthening them by using key strategies to meet the methodological challenges of studying peer groups and teacher involvement. As the target outcome, we focused on changes in sixth-graders' engagement over the school year, since middle school marks a time when peer relationships normatively increase in importance and the quality of teacher–student relationships typically declines (Wigfield et al., 2015).

Consistent with previous research, we expected to find joint effects that were both cumulative and contextualized. We focused on one specific pattern of contextualized effects, referred to as differential susceptibility, in which student receptiveness to peer group influences is more or less pronounced, depending on the quality of their involvement with teachers. Following SDT and SEF, we reasoned that, if after the transition to middle school, students are not able to establish warm and supportive relationships with teachers, they might become less adult oriented and more open to peer group influences (Davidson, Gest, & Welsh, 2010), thus amplifying the impact of peers. If so, then low teacher involvement during this developmental period could render students more susceptible to the impact of their peer groups, which would be especially problematic for students who hang out with disaffected peers. In contrast, high teacher involvement might be able to protect students from some of the motivational costs of belonging to disaffected peer groups.

To conduct this investigation, we relied on a data set that contained all of the elements needed to examine joint effects, that is, a data set that incorporated different sources of information about each of the key constructs, in this case, information about peer groups derived from multiple peer observers, ratings of student engagement from teachers, and students' ratings of their experiences of the involvement provided by their teacher (Kindermann, 2007). Although peer group contributions to student engagement, through processes of selection and socialization, have been documented in this data set, no previous attempts have been made to determine whether the magnitude of these effects differs for students who experience differing levels of teacher involvement. In some ways, the current study may be seen as encouragement to researchers who have previously examined the contributions of peers or teachers separately, to revisit their data sets to see if information about the other social partner is available, and so would allow a more ecologically oriented examination of their joint effects on student engagement, motivation, or adjustment.

We investigated patterns of joint influence in three steps. First, we examined the possibility of cumulative effects, in which teachers and peer groups make largely separate and additive contributions to students' developing academic engagement. To test this model, we first replicated the general finding that teacher involvement and peer group profiles of engagement each positively predicts changes in student engagement individually, and then examined whether they make additive contributions. We expected that peer groups would make a unique contribution, over and above the

contribution of teacher involvement. Second, we investigated the possibility of contextualized effects. Consistent with the notion of differential susceptibility, we expected that peer groups would play a more prominent role in predicting changes in students' engagement when teachers were less involved. Based on research suggesting that peer groups can socialize toward engagement or toward disaffection (Kindermann, 2007; A. M. Ryan, 2001; Wang & Eccles, 2012), we expected that these more pronounced peer-group contributions would be positive or negative depending on the profile of engagement versus disaffection characterizing each child's peer group, with students who affiliated with engaged peers groups showing increases in engagement and those affiliating with disaffected groups showing declines over the school year.

Third, we explored patterns that included both cumulative and contextualized effects, using latent profile analysis to identify groups of students who showed different combinations of teacher involvement and peer-group engagement. We then examined whether these clusters of students showed different patterns of change in their academic engagement over the year. We expected to see two specific patterns. First, we predicted cumulative effects: It was expected that neither having an involved teacher alone nor affiliating with engaged peers alone would be sufficient to foster optimal levels of student engagement. To optimize engagement, students would likely require both involved teachers and engaged peer groups. If so, then students with the highest levels of engagement over the year would be those who both affiliated with engaged peers and experienced high levels of teacher involvement, whereas the steepest declines would be found among students who not only affiliated with disaffected peers, but also experienced their teachers as uninvolved. Second, we also expected contextualized effects, such that high teacher involvement would protect children from some of the motivational costs of affiliating with disaffected peer groups (Sabol & Pianta, 2012), and by the same token, connections with engaged peers would buffer students from the motivational costs of experiencing uninvolved teachers.

Method

For this study, Kindermann's (2007) dataset was reanalyzed. Of 366 sixth-grade students (ages 11–13) enrolled at the sole middle school (Grades 6 through 8) in a small rural/suburban town in the United States, 340 (93%) participated; all of them had been participants in a longitudinal study since third grade. Most students identified themselves as Caucasian, with less than 5% identifying themselves as non-White, and were predominately from working- to middle-class families (87% of the adult population had at least a high school degree). The number of male and female participants was roughly equivalent (48% female).

The middle school these sixth graders attended was organized around homeroom classes: Students were assigned to these structured 20-min first-period classes for the whole year. This arrangement was explicitly designed to provide homeroom teachers with the opportunity to get to know their students by checking in and interacting with them every day. Although homeroom teachers taught varying subjects (and so saw most of their students again in content classes), they were expected to serve as supports for their homeroom students and as designated liaisons to other teachers if students experienced academic or behavioral problems. All 13 of the sixth-grade homeroom teachers participated in the current study. They provided information about the students in their homeroom classes, and indicated that they knew their students very well and were familiar with their academic problems and progress. Questionnaires were administered to students in class by trained interviewers; items were read aloud by one interviewer, while a second interviewer monitored the classroom to answer individual students' questions. Teachers were not present in the classroom, and typically completed their questionnaires during this time.

Students' Academic Engagement

Students' academic engagement was assessed using a 14-item Likert-type scale measuring teachers' perceptions of students' engagement in academic activities (Wellborn, 1992). These measures are not intended to measure engagement in a single classroom, but in classrooms in general. The scale assesses students' behavioral engagement (e.g., "This student works as hard as he/she can") and emotional engagement (e.g., "In my class, this student appears happy"). Prior studies on fourth through seventh graders have shown moderate to strong intercorrelations between the components ($r = .72$, $n = 1{,}018$; Skinner, Kindermann, & Furrer, 2009) and indicated that they form an internally consistent indicator of engagement ($\alpha = .90$, $n = 1{,}018$). Teacher reports of engagement have been found to be stable over time ($r = .73$, $p < .001$, $n = 144$; Wellborn, 1992; $r = .78$, $p < .001$, $n = 1{,}018$; Skinner et al., 2008) and moderately correlated with academic achievement in the expected direction ($r = .40$ with math achievement, $r = .58$ with reading achievement; Skinner & Belmont, 1993; Skinner et al., 1990).

Teacher perceptions of student engagement were obtained at two time points during the school year, first in October and then again in May. At the first time point, homeroom teachers reported on 318 students (93% of the consenting students; 87% of the population). At the second time point, homeroom teachers reported on 322 students. Missing data and differences in sample size at the two measurement points are due to a combination of student attrition and new students entering the school. Three hundred students had teacher reports at both time points.

Naturally Occurring Peer Groups

In October, students reported on naturally occurring peer groups using SCM (Cairns et al., 1985). In SCM, participants serve as "expert observers," reporting on whom they frequently see "hanging around" together while at or away from school. Students were provided with a form containing space for observations of up to 20 groups, each group having space for up to 20 members. Of the 280 participating students (77% of the sample; 56% female), none exhausted the space provided. Students were encouraged to consider all students in their entire school, regardless of grade level, as well as peers from outside the school. They were asked to list as many groups as they could from free recall, and were instructed to include dyadic groups as well as their own groups. Students could be nominated as being members of many separate groups at the same time so that multiple and overlapping groups were retained.

Peer groups were identified by first arranging students' reports of groups in a co-occurrence matrix, indicating the frequency with which each student was observed in interactions with each other student. Binomial z scores were calculated for each co-occurrence

in the matrix, and a 1% significance level was used to determine whether a student was more likely to be nominated as being in a group with each other student than could be expected by chance (for details, see Kindermann, 2007). In order to guard against self-enhancement biases, significant connections that were based on one single observation were not accepted, as in almost all cases these were children's own self-nominations. Not counting errors of omission (e.g., that most girls do not report most boys' peer groups), there was high consensus about group connections ($\kappa = .88$).

Three key indices of the characteristics of peer group networks were calculated. The number of members, excluding the focal student, who were identified in each student's peer group was used as a measure of group size. The percentage of peers maintained as group members from fall to spring was taken as an indicator of peer group stability. Finally, peer group profiles of engagement were calculated by averaging the teacher-rated engagement scores across the members of each child's group connections.

Teacher Involvement

In October, students themselves reported on the amount of involvement experienced from their teachers by responding to 11 items (Skinner & Belmont, 1993; all items were on a 4-point scale). The scale captures three facets of teacher involvement: the extent to which students' teachers showed affection (three items; e.g., "My teacher really cares about me"), the extent of availability (three items; e.g., "My teacher is always there for me"), and the extent of dependability (five items; e.g., "I can rely on my teacher to be there when I need him/her"). Because the students had been involved in the longitudinal study from third grade onward, teacher involvement items were worded so that they referred to a single teacher. Thus, the items are used as a proxy for students' experiences of general teacher involvement. Previous work has found that these measures have high internal consistency ($\alpha = .79$, $n = 144$; Skinner & Belmont, 1993) and that their key psychometric and validity characteristics are maintained from elementary to middle school (Skinner, Kindermann, & Furrer, 2009; Skinner et al., 1998).

Results

Descriptive statistics and correlations can be found in Table 1. In all analyses, a full information maximum likelihood method was used to estimate missing data. Overall, students showed moderate levels of engagement in both fall ($M = 3.07$, $SD = .57$) and spring ($M = 3.07$, $SD = .61$), with relatively high stability between time points ($r = .75$, $p < .001$). On average, members of students' peer groups were moderately engaged in fall ($M = 3.09$, $SD = .34$), with larger groups showing a tendency toward higher engagement ($r = .25$, $p < .001$). Peer groups were modest in size ($M = 4.81$, $SD = 3.99$), and relatively stable across the school year, with just about half of students' affiliations in fall continuing into spring ($M = .46$, $SD = .33$). Finally, while students, on average, rated their teachers as being fairly involved ($M = 3.01$, $SD = .52$), students who experienced their teachers as more involved were more engaged themselves both in fall ($r = .34$, $p < .001$) and in spring ($r = .40$, $p < .001$), and tended to be affiliated with peers who were more engaged in fall ($r = .20$, $p < .001$).

Cumulative Effects: Do Peer Groups Contribute to Engagement Over and Above the Effects of Teachers?

Cumulative effects of teachers and peers were examined in two steps. First, in analyses of each potential contributor separately, peer group engagement scores in fall were found to predict changes in students' engagement from fall to spring ($\beta = .11$, $p < .05$), $\chi^2(36) = 60.250$, $p = .007$; minimum discrepancy divided by degrees of freedom (CMIN/DF) = 1.674, comparative fit index (CFI) = .988, root-mean-square error of approximation (RMSEA) = .043, as did teacher involvement ($\beta = .15$, $p < .01$), $\chi^2(39) = 51.333$, $p = .089$; CMIN/DF = 1.316, CFI = .993, RMSEA = .029. Both models controlled for sex, peer-group stability, and peer-group size. In a second set of analyses, the contributions of peer-group engagement and teacher involvement were modeled simultaneously, again controlling for peer-group size and stability, and sex. This model fit the data well, $\chi^2(66) = 101.358$, $p = .003$, CMIN/DF = 1.536, CFI = .986, RMSEA = .038, and indicated that peer group engagement in fall predicted changes in students' engagement across the school year ($\beta = .10$, $p < .05$), over and above the contribution of teacher involvement ($\beta = .15$, $p < .01$).

Contextualized Effects: Do the Effects of Peers Differ for Students With Different Levels of Teacher Involvement?

In order to investigate differential susceptibility, in which students' experiences of teacher involvement can magnify or reduce

Table 1
Construct Means, Standard Deviations, and Correlations

Variable	1	2	3	4	5	6	7	*M*	*SD*
1. Student engagement fall	—							3.07	.57
2. Student engagement spring	.75***	—						3.07	.61
3. Peer engagement fall	.42***	.39***	—					3.09	.34
4. Teacher involvement	.34***	.40***	.20***	—				3.01	.52
5. Sex	.16**	.21***	.19***	.29***	—			1.47	.50
6. Group stability	.17*	.17*	.03	.10†	.20***	—		0.46	.33
7. Group size	.19***	.08	.25***	.10†	.27***	.13*	—	4.81	3.99

Note. $N = 366$.
† $p < .07$. * $p < .05$. ** $p < .01$. *** $p < .001$.

the impact of their peer groups on academic engagement, data were analyzed using two complementary modeling strategies. First, a latent moderated model was used to examine whether peer group effects on changes in students' engagement depended on teacher involvement. Then, to aid the interpretation of the interaction, a multiple-group model examined whether peer-group effects on students' academic engagement were significantly different between students who experienced highly involved teachers and students who experienced less-involved teachers. Both modeling strategies used AMOS 19 (Arbuckle, 2010).

The moderated model (see Figure 1), which was tested using the orthoganilizing procedure described by Little, Bovaird, and Widaman (2006), showed good fit to the data, $\chi^2(203) = 188.302$, $p = .763$; CMIN/DF = .928, CFI = 1.000, RMSEA = .000, confidence interval (CI) [.000, .017]. As indicated by the model, the interaction between peer group engagement and teacher involvement was a significant predictor of changes in students' engagement from fall to spring ($\beta = -.15$, $p < .01$), suggesting that peer group contributions to changes in students' academic engagement were dependent on students' experiences of involvement from

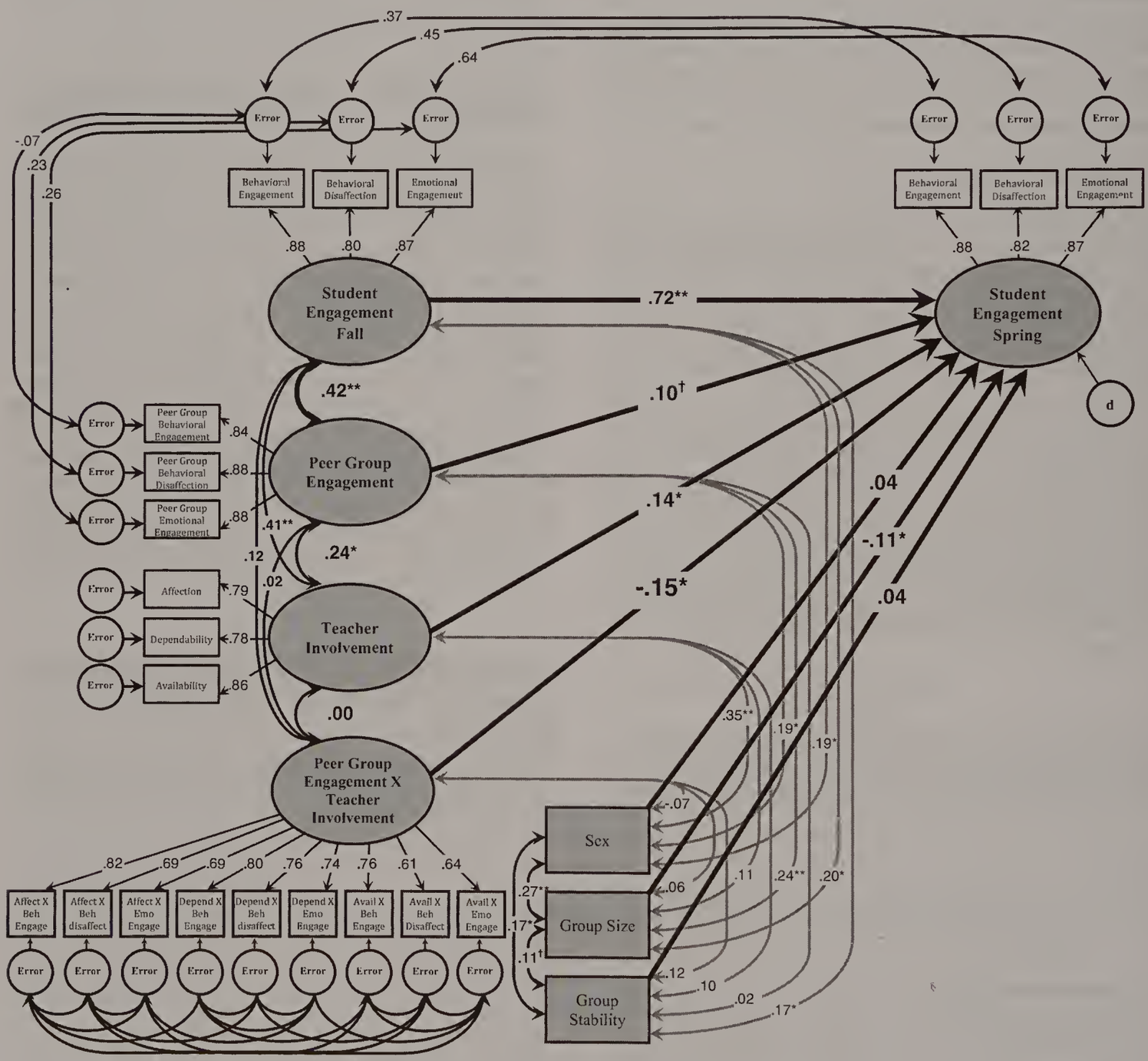


Figure 1. The interplay between peer and teacher influences on changes in students' academic engagement; $\chi^2(203) = 188.302$, $p = .763$; minimum discrepancy divided by degrees of freedom = .928; comparative fit index = 1.000; root-mean-square error of approximation = .000; 90% confidence interval, [.000, .017]. * $p < .01$. ** $p < .001$. † $p < .07$.

their teachers. The negative interaction parameter indicates that lower teacher involvement was associated with increased peer effects.

Multiple-group structural equation modeling (SEM) analyses further explored whether differences in the extent to which peer groups contributed to students' engagement could be found between students who experienced highly involved teachers versus students who experienced their teachers as less involved. Using a tertile split, two groups of nearly equivalent size were identified: students who perceived their teachers as most involved (n = 129, mean involvement = 3.54), and students who perceived their teachers as least involved (n = 127, mean involvement = 2.47; 100 students in the middle range were omitted). Compared to students who experienced teachers as least involved, students of highly involved teachers were more engaged both in fall and spring, and affiliated with more engaged peers (see Table 2).

To test for differences in peer-group effects between these groups, a three-step model invariance procedure was used (Kline, 2011; Tabachnick & Fidell, 2007). First, a configural model (see Figure 2) was fit to the data, freely and simultaneously estimating model parameters for both groups. This model showed good fit to the data, $\chi^2(72) = 96.454$, $p = .029$; CMIN/DF = 1.340; CFI = .982; RMSEA = .038; 90% CI [.013, .057]. Despite similarity between the two groups in terms of stability of individual engagement, peer group contributions to changes in student engagement were greater for students who perceived their teachers as less involved ($\beta = .30$, $p < .001$). By comparison, results indicated that no significant peer group effects were found among students who perceived their teachers as most involved ($\beta = -.05$, $p > .05$).

To test the significance of this difference, cross-group equality constraints were imposed on the model, beginning with the factor loadings. Constraining the measurement portion of the model did not lead to significant reductions in model fit, $\Delta\chi^2(6) = 5.548$, $p > .05$, indicating measurement equivalence between the two groups. The model fit the data well, $\chi^2(78) = 96.180$, $p = .080$; CMIN/DF = 1.233; CFI = .987; RMSEA = .030; 90% CI [.000, .049]. In a final step, the two-group model was estimated with an additional cross-group equality constraint imposed on the model parameter representing peer-group effects on changes in students' engagement. Model fit remained good, $\chi^2(79) = 102.337$, $p = .040$; CMIN/DF = 1.295; CFI = .984; RMSEA = .034; 90% CI [.008, .052]; however, the imposition of this constraint lead to a significant reduction in model fit, $\Delta\chi^2(1) = 6.157$, $p < .05$, confirming the expectation that this parameter of the model should differ between groups. These results complement findings from the latent moderated model and support a contextualized view of peer influence on student engagement, suggesting that peer-group contributions to students' engagement were greater among students who experienced their teachers as less involved.

Cumulative and Contextualized Effects: Do Students With Different Configurations of Peer-Group Engagement and Teacher Involvement Show Differential Change in Engagement?

To explore cumulative and contextualized effects, that is, to examine whether the joint effects of teachers and peer groups can be additive and compensatory, data were analyzed using a person-centered approach. Specifically, groups of students were identified who had different combinations of peer and teacher contexts, and these groups were compared to see whether they differed in the way their engagement changed across the school year. For these analyses, types of students were identified from the entire sample using latent profile analysis (LPA) modeling using MPLUS (version 7.2; Muthén & Muthén, 2012). Four separate LPA models were tested, with two, three, four, or five profiles specified. All LPA model solutions were stopped at 60,000 iterations, and relative fit was assessed by comparing the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and the Sample Size-Adjusted Bayesian Information Criterion (Adj. BIC) for each model (see Table 3). With each of these assessments of fit, lower values indicate better model fit (Nylund, Asparouhov, & Muthén, 2007). In addition to assessing relative model fit, model preference was also based on the presence of adequately sized profile groupings, and whether groups fit well with theoretical expectations. While the five-profile model showed the best fit (as assessed by AIC and Adj. BIC), one of the profile groupings identified was inadequately sized ($n < 10$), and substantively indistinguishable from another profile grouping; thus, profiles identified using this model were not used. Profiles based on the four-profile model were chosen for use in subsequent analyses, as this model showed the best fit (in comparison to all models but the five-profile model), produced adequately sized profile groupings, and aligned well with theoretical expectations for the variety of configurations represented. This model identified the following types of students: members of engaged peer groups (M = 3.35, SD = .25) who experienced high teacher involvement (M = 3.52, SD = .30; n = 132), members of engaged peer groups (M = 3.28, SD = .24) who experienced low teacher involvement (M = 2.56, SD = .32; n = 107), members of disaffected peer groups (M = 2.74, SD = .20)

Table 2
Mean Level Differences Between Students With High or Low Teacher Involvement

Variable	High teacher involvement		Low teacher involvement				
	M	*SD*	*M*	*SD*	M_{diff}	*t*	*p*
Student engagement fall	3.29	.53	2.85	.58	−.44	−6.32	<.001
Student engagement spring	3.32	.55	2.76	.61	−.55	−7.62	<.001
Peer engagement fall	3.18	.33	3.02	.35	−.16	−3.83	<.001
Peer group size	5.57	3.99	4.94	4.20	−.64	−1.24	>.05
Peer stability	0.50	.34	0.44	.33	−.06	−1.42	>.05

Note. High teacher involvement group, n = 129; low teacher involvement group, n = 127.

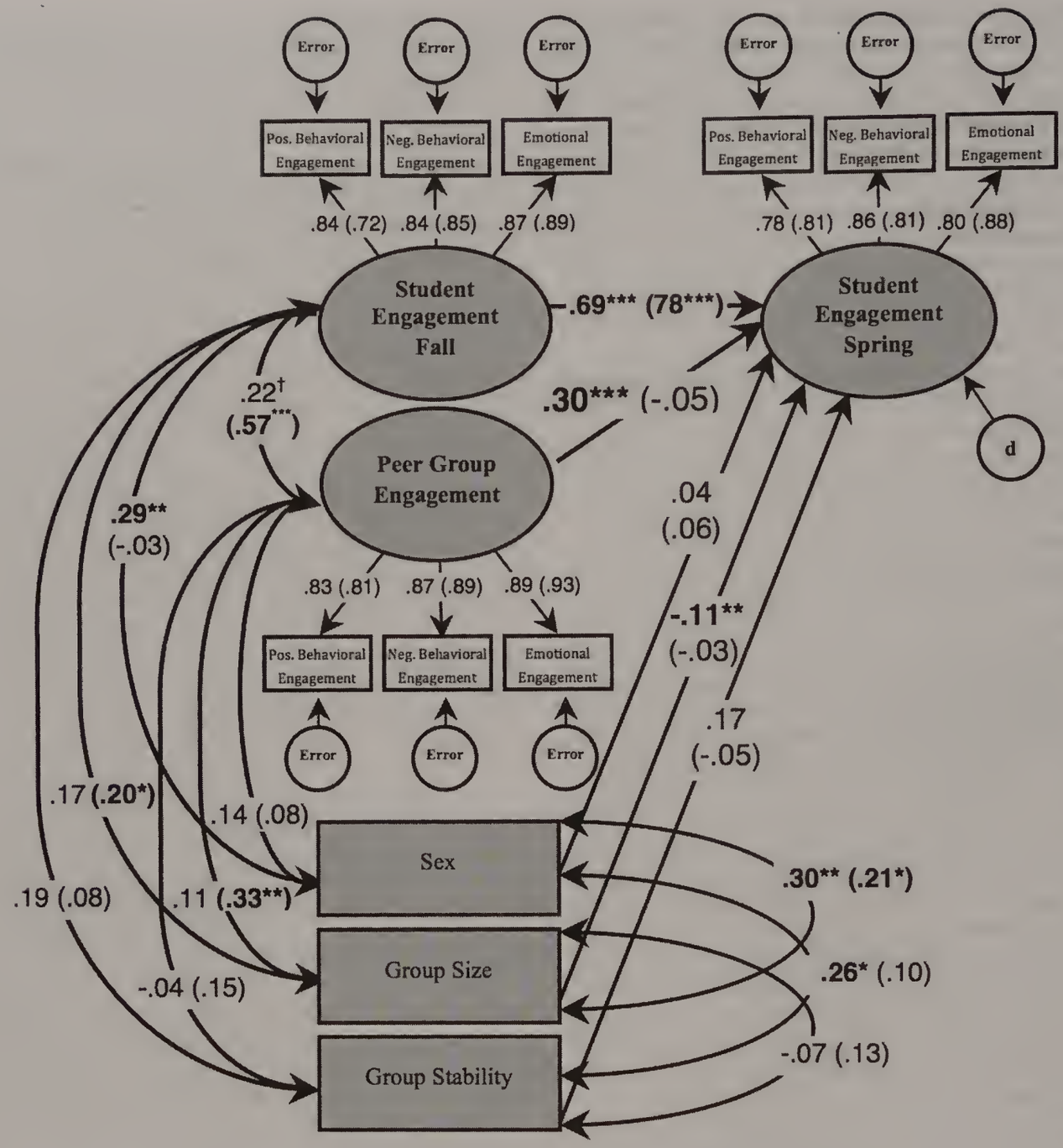


Figure 2. Comparison of peer influence between students who perceive teachers as highly involved and students who perceive teachers as least involved; $\chi^2(72) = 90.632$, $p = .068$; minimum discrepancy divided by degrees of freedom = 1.259; comparative fit index = .987; root-mean-square error of approximation = .032; 90% confidence interval, [.000, .051]. The model parameters for students who experienced their teacher as highly involved appear in parentheses. Error correlations have been omitted from the figure for clarity. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$. $^{\dagger}p < .07$.

who experienced high teacher involvement ($M = 3.08$, $SD = .33$; $n = 94$), and members of disaffected peer groups ($M = 2.50$, $SD = .31$) who experienced low teacher involvement ($M = 2.21$, $SD = .41$; $n = 33$).

Changes in engagement across the school year are shown in Figure 3 for each of the four different types of students identified by LPA. An analysis of covariance (ANCOVA) showed mean level differences in student engagement among the four types of students in both fall, $F(3, 359) = 15.66$, $p < .001$, and spring, $F(3, 359) = 21.94$, $p < .001$, accounting for the covariates of sex, peer group size, and group stability. The strength of the relationship between cluster membership and student engagement in both fall ($\omega^2 = .10$) and spring ($\omega^2 = .14$) was moderate, accounting for the covariates. Post hoc comparisons examined pairwise differences between the groups. As expected, the most engaged students were those who both affiliated with highly engaged peers and also experienced teachers as highly involved; these students showed the highest levels of engagement in both fall ($M = 3.32$, $SD = .51$)

Table 3
Latent Profile Analysis Model Fit Results

Model	Statistical criteria		
	AIC	BIC	Adj. BIC
Two profile	1102.421	1129.739	1107.531
Three profile	1053.001	1095.93	1061.032
Four profile	**1029.088**	**1087.627**	**1040.038**
Five profile	1024.618	1098.768	1038.489

Note. AIC = Akaike Information Criteria; BIC = Bayesian Information Criteria; Adj. BIC = sample-size adjusted Bayesian Information Criteria, $n^* = (n + 2)/24$. Lower AIC, BIC, and Adj. BIC values indicate better fit. Indices of fit for the chosen model appear in boldface type. The five-profile model was removed from consideration because it identified profiles of inadequate size ($n < 10$).

and spring ($M = 3.34$, $SD = .55$). They were more engaged than students who experienced only favorable teacher contexts (M difference = .33, $p < .001$, in fall; mean difference = .27, $p < .001$, in spring), as well as students who experienced only favorable peer contexts (mean difference = .31, $p < .001$, in fall; mean difference = .33, $p < .001$, in spring). Conversely, the most disaffected students were those who both affiliated with disaffected peers and experienced teachers as least involved; these students showed the lowest levels of engagement in both fall ($M = 2.62$, $SD = .62$) and spring ($M = 2.41$, $SD = .63$). They were less engaged than students who experienced only favorable teacher contexts (mean difference = $-.32$, $p < .01$, in fall; mean difference = $-.61$, $p < .001$, in spring), as well as students who experienced only favorable peer contexts (mean difference = $-.34$, $p < .01$, in fall; mean difference = $-.54$, $p < .001$, in spring). Together these results support an additive model of the joint contributions of peer groups and teachers to students' engagement: Although it was better to have either an engaged peer group or an involved teacher than having neither, for students to have the highest levels of engagement, support from both peers and teachers was needed.

Results from a repeated-measures ANCOVA also showed significant differences between the LPA-identified groups in how their engagement changed across the school year, $F(3, 359) = 3.93$, $p < .01$. As predicted, students who experienced teacher and peer group contexts that were both favorable (i.e., having involved teachers and affiliating with engaged peer groups) fared best over time, showing high and stable engagement across the year ($\Delta M = .01$), $t(131) = .40$, $p > .05$. In contrast, students with the least favorable contexts (i.e., who experienced the least involvement from their teachers and also affiliated with the most disaffected peers) demonstrated the steepest declines in engagement across the academic year ($\Delta M = -.21$), $t(40) = -2.35$, $p < .05$. At the same time, evidence for partially compensatory effects was also found. Students who affiliated with disaffected peers, but who viewed their teachers as more involved evinced moderate levels of academic engagement (showing higher levels than the most disaffected group but lower levels than the most engaged group) that increased marginally ($\Delta M = .09$) from fall to spring, $t(93) = 1.91$, $p < .06$. This suggests that an involved teacher can offset some of the motivational costs of affiliating with disengaged peers. Similarly, students who viewed their teachers as less involved, but nevertheless affiliated with engaged peers also showed moderate levels of academic engagement that remained stable over the year ($\Delta M = -.01$), $t(106) = -.38$, *ns*, suggesting that academically enthusiastic peer group members can protect against some of the motivational costs of experiencing teachers as unsupportive.

Discussion

Urie Bronfenbrenner (1976) wrote, "the ecology of education is not and cannot be confined solely to conditions and events occurring within a single setting, such as home, school, peer group, workplace, etc.; equal emphasis must be given to relations obtaining between settings" (p. 12). If the notion of "contextualization" can be applied not only to settings, but also to interactions with social partners, this suggests that the nature of the interactions between two people in a given setting may be best understood in the context of the other interactions those two people have experienced in that setting. In that spirit, this study sought to contribute to an emerging body of work focused on the joint effects of teachers and peers by examining how these subsystems of the school social ecology work together, both independently and interdependently, to shape students' motivation to engage in learning activities.

More specifically, we tested whether models of cumulative (additive) and contextualized (interactive) joint effects could explain changes in students' engagement from fall to spring of their first year in middle school, when peer influences are on the rise and the quality of students' relationships with their teachers typically declines. Evidence was found for both kinds of effects. Consistent with prior studies (Kindermann, 2007; A. M. Ryan, 2001; Klem & Connell, 2004; Wentzel, 1997), results from SEM models testing peer and teacher effects separately indicated that both peer group engagement and teacher involvement individually predict changes in student engagement over the school year. When tested simultaneously, peer groups were found to make a unique contribution to changes in students' engagement, over and above the substantial contribution of teacher involvement, suggesting that peers and teachers contribute uniquely to students' engagement, and that their effects may be cumulative.

At the same time, findings suggested that the effects of peer groups are also contextualized. Interactions between peer groups' engagement and teachers' involvement were significant as predictors of changes in students' own engagement, indicating that the motivational contribution of peer groups was magnified or reduced depending on students' experiences of involvement from their teachers. Consistent with the notion of differential susceptibility, tests of multigroup models showed that peer groups were significantly stronger as predictors of changes in engagement among students who perceived their teachers as less involved, with these more pronounced peer group effects associated with positive or negative consequences for engagement depending on the motivational composition of each child's peer group.

Person-centered analyses likewise revealed support for both cumulative and contextualized models. On the one hand, joint effects of peer groups and teachers were clearly cumulative. Neither teacher involvement nor peer-group engagement alone were sufficient to foster the highest levels of student engagement; and declines in engagement were steepest for students who both affiliated with disaffected peers and reported lower levels of teacher involvement. On the other hand, results also suggested that teacher

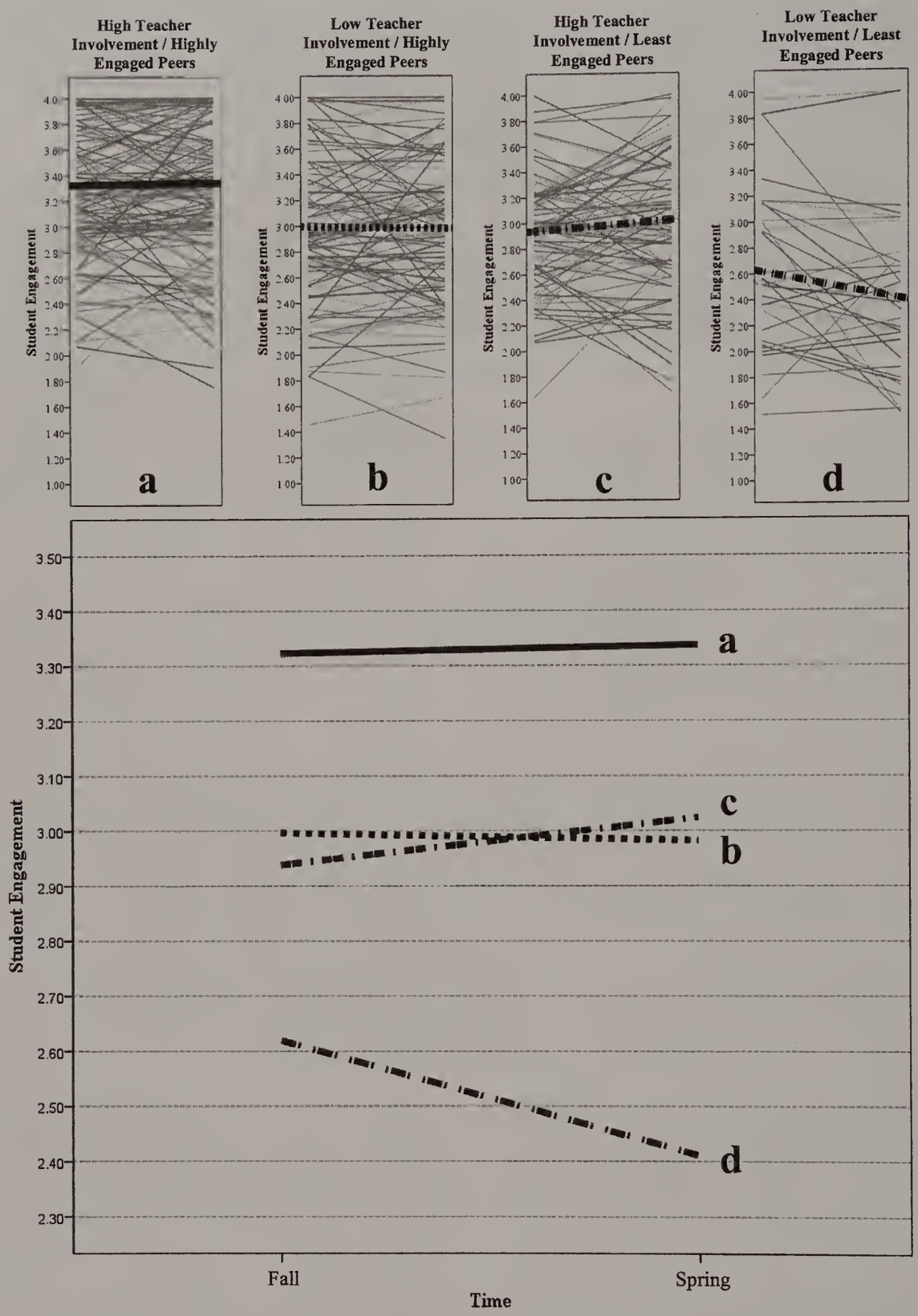


Figure 3. Differential growth of students' academic engagement based on students' combined experiences of teacher involvement and peer group engagement. Nonsignificant engagement changes in academic engagement are shown for students who (a) affiliate with highly engaged peer and experience high teacher involvement (n = 132), and who (b) affiliate with highly engaged peers and experience low teacher involvement (n = 107). Gains in engagement shown for students who (c) affiliate with least-engaged peers and experience high teacher involvement (n = 94; average increase marginally significant at $p < .06$). Significant decreases in engagement shown for students who (d) affiliate with least-engaged peers and experience low teacher involvement (n = 33; average decreases significant at $p < .05$).

involvement could partially buffer students from the motivational costs of belonging to disaffected peer groups: Students who affiliated with disengaged peers, but still experienced teachers as involved, showed moderate levels of engagement and made marginally significant gains in engagement across the school year. Perhaps surprisingly, these configural analyses also suggested that peer groups can dampen the effects of low involvement from teachers: Students who experienced their teachers as less involved, but who nevertheless affiliated with more engaged peers, also showed moderate levels of engagement and were able to maintain their engagement over the school year. Perhaps positive connections with peers have the potential, at least in the short run of a school year, to buffer some of the motivational costs otherwise associated with perceived lack of teacher support.

Limitations and Future Directions

It is important to consider the shortcomings of the current study, in terms of design and measurement, when interpreting its findings and making suggestions for further investigation of joint effects. In terms of design, the study is limited in that it focused on changes in engagement across only two time points within a single year. This made it impossible to follow the joint effects of teachers and peers across subsequent school years, as students encountered new teachers and joined new peer groups. Longer time frames with more measurement points would be useful if future studies aim to examine cumulative long-term effects or to explore meditational models, reciprocal effects, or growth curves. Previous studies of joint effects suggest that all of these processes are important targets (e.g., Wang & Eccles, 2012).

In terms of measurement, the strategies used in the current study to assess student engagement and teacher involvement were limited in their ability to map the complex social world of teachers and peers during middle school. The use of multiple, independent reporters can be viewed as a strength of the study, but the decision to ask students about the involvement of "my teacher" as a proxy for the involvement of teachers in general may not fully capture the range of teacher interactions that students experience in middle school settings, where they typically encounter many teachers over the day. Although this strategy is common among researchers who assume that it is students' perceptions of experiences that are key to their engagement (e.g., DeLaet et al., 2015; Furrer & Skinner, 2003; Skinner & Belmont, 1993), it is an empirical question whether findings from the current study will replicate in research utilizing aggregate indicators that, for example, are based on average involvement scores from every teacher with whom a student interacts. By the same token, reliance on homeroom teachers as the sole reporters of students' engagement, even teachers who indicated that they knew students well, may fall short of capturing the range of engagement that students exhibit from class to class over the day. Although teacher reports of student engagement likely have advantages over self-reports (which were used in the majority of studies targeting joint teacher–peer effects), it is an empirical question whether the pattern of results found in the current study will replicate in research focused on other measures of engagement, such as classroom observations or aggregates that combine engagement ratings from multiple teachers.

Finally, the current study is limited in that it did not supplement longitudinal correlational findings by explicitly incorporating markers of potential mechanisms of joint influence. As previously discussed, a variety of mechanisms have been documented through which teachers influence student engagement and motivation (Sabol & Pianta, 2012; Wentzel, 2009b) and studies are increasingly identifying pathways of peer influence, which seem to be both cognitive (for a review, see Brechwald & Prinstein, 2011) and behavioral (Kandel, 1985; Sage & Kindermann, 1999). Such evidence bolsters the current correlational findings, but causal interpretations would be strengthened by future studies that include measures of possible mechanisms to explain joint teacher–peer effects, and then test their viability using meditational analyses. Combined with experimental studies, such findings would help rule out alternative third-variable explanations that are otherwise plausible. For example, students' behavioral problems may underlie both declines in their academic motivation and in the quality of their relationships with teachers (e.g., Wang & Fredricks, 2014) and peers (Davidson et al., 2010). Future studies that directly examine potential mechanisms would begin to identify the (perhaps multiple) pathways through which peers and teachers jointly influence students' academic engagement.

Implications for Future Research

The current study is consistent with previous research examining the joint effects of peer and teacher relationships on the development of students' academic functioning, but also makes several key contributions to this growing area of study. First, findings suggest that, in addition to the features of peers already identified in other studies of joint teacher–peer effects, it would be useful to add peer groups, or more specifically, the motivational composition of peer groups, as another peer attribute that plays a role in students' engagement over and above that of teachers, and whose effects seem to be contextualized by teacher involvement. Second, findings from the current investigation corroborate the notion that using distinct sources of information about the three players in processes of joint effects (namely, peers, teachers, and student engagement) may make it easier to discern certain forms of contextualized effects. Third, it underscores some of the methodological strategies, like SCM, that may be useful in capturing the active ingredients in peer groups, and encourages researchers to consider reanalyzing their data sets, if they contain all the elements needed to meaningfully test for cumulative and contextualized joint effects.

Finally, the current study highlights the value of using an ecological perspective, as well as motivational and developmental theories like SDT and SEF, to frame expectations about joint influences. Ecological perspectives provide a larger framework within which to consider the influences of peer and teacher subsystems, and suggest conceptual terms, like cumulative and contextualized effects, to supplement researchers' reliance on statistical terms like additive and interactive. They open up a range of other kinds of contextualized effects and point to other subsystems, such as friendship networks, or family and neighborhood subsystems, to which these ideas could usefully be extended. More specific theories, like SDT, suggest that future work examining potential mechanisms should include students' sense of relatedness to peers and teachers, and examine whether they mediate the effects of peer-group engagement and teacher involvement on changes in student engagement. SEF also suggests that particular

kinds of contextualized effects, in which low teacher involvement contributes to differential susceptibility to peer groups, may represent an emergent developmental phenomenon, that only appears after the transition to middle school, when environmental shifts make it more difficult for students to connect with teachers (Eccles & Roeser, 2009). Such conceptual considerations may be helpful to future studies in guiding the selection of peer characteristics and in making predictions about how and why their effects might (or might not) be contextualized by students' relationships with their teachers.

Models of joint teacher–peer effects. Future research on joint effects may also benefit from greater discussion of the different ways in which influences from peers and teachers can work together. The notion of differential susceptibility provides one hypothesis, in which the lack of close relationships with teachers renders students more open to peer group influences (Sabol & Pianta, 2012), but alternative models that posit other kinds of contextualized effects, like compensatory or synergistic effects, could also be fruitful. It is important to note that these alternative models are not necessarily mutually exclusive; some are complementary. For example, in the present study, we found evidence that joint effects are both cumulative (i.e., additive) and contextualized (in this case, amplifying and dampening susceptibility). In other words, both teachers and peers are clearly important, in that support from one partner cannot fully substitute for poor relationships with the other, but each partner can still buffer or protect students from the worst motivational consequences of low levels of support from the other.

Of course, some models of contextualized effects are incompatible with cumulative models. For example, fully compensatory models, in which high levels of support from either partner are sufficient to produce the best outcomes, indicate that effects are not additive, instead they are substitutive—either one is a sufficient condition for the outcome. Other kinds of contextualized models also rule out additive effects, such as multiplicative threshold models in which some minimal level of support from one partner is required if the other partner is to have an impact. For example, if relationships with teachers are bad enough, it may be that no amount of peer encouragement can reignite students' engagement. Or if a student's peer group is sufficiently disaffected, he or she may no longer respond to a teacher's involvement, support, or reassurance. These models suggest completely contextualized effects, where if one relationship is unfavorable enough, it can actually cancel the impact of the other social partner.

One strategy for discerning contextualized effects, used in the current study as well as several previous ones (Davidson et al., 2010; Furrer & Skinner, 2003; Raufelder et al., 2013), is the use of pattern-oriented or person-centered approaches that identify subgroups of students who inhabit qualitatively different peer–teacher niches, and comparing them on target outcomes, such as changes in motivational or academic functioning. The more that research on joint effects moves away from statistical models of interactive effects and toward conceptual models of contextualized effects, the wider the array of methodological strategies that can be brought to bear. In this regard, researchers may wish to take advantage of conceptual models and statistical techniques applied in work on developmental psychopathology (Luthar, Cicchetti, & Becker, 2000) and environmental reactivity (e.g., Moore & Depue, 2016), where researchers have considered hypotheses that include protective or buffering effects, immunization, thresholds, cumulative risk, diathesis stress, and other forms of differential vulnerability or susceptibility to the environment (e.g., Ellis et al., 2011).

Reciprocal effects of student engagement on teachers and peers. As research increasingly focuses on social dynamics among multiple partners in the classroom, studies may also be expanded to include a consideration of reciprocal effects, in which students' own engagement feeds back to shape the supports they receive from teachers and peers. And just as feedforward effects have been found to be contextualized, it is possible that reciprocal processes involving both social partners may also interact with each other. Engagement versus disaffection may turn out to be markers for whether students receive a double dose of motivational support or discouragement. That is, students who are highly engaged not only receive more involvement (as well as other forms of support) from teachers, but they also have access to more engaged groups of peers, whereas more disaffected students typically experience their teachers as withdrawing their support and becoming more controlling over time, while at the same time their peer connections are largely confined to other disaffected students (Kiuru et al., 2015; Nurmi & Kiuru, 2015; Skinner & Belmont, 1993). Future studies that examine joint effects over longer periods of time could explore whether such reciprocal feedback processes amplify the feedforward effects suggested by findings from the current study, potentially contributing to virtuous and vicious cycles that shape the development of student engagement and motivation over multiple school years.

Implications for Practice

Although the research base is too thin at the current time to allow for any definitive recommendations, findings from studies of joint effects suggest three possibilities for educators and researchers to consider in their efforts to refine practices and strengthen interventions designed to promote students' engagement and motivational development. First, it seems likely that interventions targeting either teachers or peers many exert their effects through two pathways. Cumulative and contextualized joint effects, such as the ones found in the current study, imply that improvements in connections with either partner (e.g., increasing connections with teachers or with engaged peers) should not only exert positive effects on engagement directly, but should also exert positive effects indirectly, by mitigating the worst impacts of problems with the other partner (e.g., low-quality relationships with teachers or connections to disaffected peers). At the same time, however, studies of joint effects also suggest that interventions targeting only one social partner will not be sufficient to optimize student engagement over the long term. Results from this and other studies indicating contextualized effects imply that interventions focusing on a single classroom partner will only produce optimal engagement for a subset of students, namely, those who already have positive connections with the other partner. The students most in need of support, namely, highly disaffected students, will likely improve only in response to interventions that help them establish improved relationships with both involved teachers and engaged groups of peers.

Second, if the effects of peers are indeed contextualized, then interventions designed to improve student–teacher relationships

may not only take on an added urgency, they may also benefit from an expanded focus. The sense of urgency follows from findings suggesting that low-quality teacher relationships may pose a double risk for student motivation: once because of the direct impact of unsupportive teachers, and once because poor student–teacher relationships may leave students at the mercy of peer-group influences, which are unlikely to be uniformly positive. According to studies of joint effects, it would be especially important for teachers to reach out to children and youth who affiliate with disaffected peers. If teachers can intentionally provide higher levels of involvement, such students may be protected from the worst effects of these connections. In the long run, teachers may develop strategies that are effective in bringing whole groups of disaffected peers back toward engagement, which would then allow all students access to groups of more engaged peers (Furrer, Skinner, & Pitzer, 2014). Interventions to support teachers in these challenging tasks may be able to bolster their resolve by highlighting findings from studies such as the current one that make more explicit the "invisible hand of the teacher" in peer relationships (Kindermann, 2011).

Finally, findings from the current study may lead researchers, interventionists, and educators to a renewed appreciation of engagement, not only as a malleable motivational state that protects students from risky behaviors and contributes to their academic success, but also as an energetic resource that students themselves can offer their classmates. Adding to the list of peer attributes that predict motivational development, such as peer relatedness, kindness, emotional support, and instrumental help, the current study highlights the potential impact of joint activity with peers who are behaviorally and emotionally engaged (Wentzel et al., 2010). Such interactions may help to sustain or rekindle students' own enthusiastic participation in learning activities. The more that theories and research can succeed in capturing the interplay among interaction partners in the complex social ecology of the school, the more helpful they will be to educators and interventionists dedicated to the hard work of optimizing students' engagement, motivation, and academic development.

References

Altermatt, E. R., & Pomerantz, E. M. (2005). The implications of having high-achieving versus low-achieving friends: A longitudinal analysis. *Social Development, 14,* 61–81. http://dx.doi.org/10.1111/j.1467-9507.2005.00291.x

Anderman, L. H., & Anderman, E. M. (1999). Social predictors of changes in students' achievement goal orientations. *Contemporary Educational Psychology, 24,* 21–37. http://dx.doi.org/10.1006/ceps.1998.0978

Anderson, A. R., Christenson, S. L., Sinclair, M. F., & Lehr, C. A. (2004). Check & connect: The importance of relationships for promoting engagement with school. *Journal of School Psychology, 42,* 95–113. http://dx.doi.org/10.1016/j.jsp.2004.01.002

Arbuckle, J. L. (2010). *Amos* (Version 19.0) [Computer software]. Chicago, IL: SPSS.

Berndt, T., Hawkins, J., & Jiao, Z. (1999). Influences of friends and friendships on adjustment to junior high school. *Merrill-Palmer Quarterly, 45,* 13–41.

Blondal, K. S., & Adalbjarnardottir, S. (2012). Student disengagement in relation to expected and unexpected educational pathways. *Scandinavian Journal of Educational Research, 56,* 85–100. http://dx.doi.org/10.1080/00313831.2011.568607

Brechwald, W. A., & Prinstein, M. J. (2011). Beyond homophily: A decade of advances in understanding peer influence processes. *Journal of Research on Adolescence, 21,* 166–179. http://dx.doi.org/10.1111/j.1532-7795.2010.00721.x

Bronfenbrenner, U. (1976). *The experimental ecology of education.* Paper presented at the annual meetings of the American Educational Research Association, San Francisco, CA.

Bronfenbrenner, U., & Morris, P. A. (2006). The bioecological model of human development. In R. Lerner & W. Damon (Eds.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (6th ed., pp. 793–828) Hoboken, NJ: Wiley.

Cairns, R. B., & Cairns, B. D. (1994). *Lifelines and risks: Pathways of youth in our time.* New York, NY: Cambridge University Press.

Cairns, R. B., Perrin, J. E., & Cairns, B. D. (1985). Social structure and social cognition in early adolescence: Affiliative patterns. *Journal of Early Adolescence, 5,* 339–355. http://dx.doi.org/10.1177/0272431685053007

Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy and relatedness: A motivational analysis of self-system processes. In M. Gunnar & L. A. Sroufe (Eds.), *Minnesota Symposium on Child Psychology: Vol. 23. Self-processes in development* (pp. 43–77). Chicago, IL: University of Chicago Press.

Danielsen, A. G., Wiium, N., Wilhelmsen, B. U., & Wold, B. (2010). Perceived support provided by teachers and classmates and students' self-reported academic initiative. *Journal of School Psychology, 48,* 247–267. http://dx.doi.org/10.1016/j.jsp.2010.02.002

Davidson, A. J., Gest, S. D., & Welsh, J. A. (2010). Relatedness with teachers and peers during early adolescence: An integrated variable-oriented and person-oriented approach. *Journal of School Psychology, 48,* 483–510. http://dx.doi.org/10.1016/j.jsp.2010.08.002

Deci, E. L. (1992). On the nature and function of motivational theories. *Psychological Science, 3,* 167–171. http://dx.doi.org/10.1111/j.1467-9280.1992.tb00020.x

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior.* New York, NY: Plenum Press. http://dx.doi.org/10.1007/978-1-4899-2271-7

De Laet, S., Colpin, H., Vervoort, E., Doumen, S., Van Leeuwen, K., Goossens, L., & Verschueren, K. (2015). Developmental trajectories of children's behavioral engagement in late elementary school: Both teachers and peers matter. *Developmental Psychology, 51,* 1292–1306. http://dx.doi.org/10.1037/a0039478

Eccles, J. S., Midgley, C., Wigfield, A., Buchanan, C. M., Reuman, D., Flanagan, C., & MacIver, D. M. (1993). Development during adolescence. The impact of stage-environment fit on young adolescents' experiences in schools and in families. *American Psychologist, 48,* 90–101. http://dx.doi.org/10.1037/0003-066X.48.2.90

Eccles, J. S., & Roeser, R. W. (2009). Schools, academic motivation, and stage-environment fit. In R. M. Lerner & L. Steinberg (Eds.), *Handbook of adolescent psychology: Vol. 1. Individual bases of adolescent development* (3rd ed., pp. 404–434). Hoboken, NJ: Wiley.

Ellis, B. J., Boyce, W. T., Belsky, J., Bakermans-Kranenburg, M. J., & van Ijzendoorn, M. H. (2011). Differential susceptibility to the environment: An evolutionary–neurodevelopmental theory. *Development and Psychopathology, 23,* 7–28. http://dx.doi.org/10.1017/S0954579410000611

Fall, A. M., & Roberts, G. (2012). High school dropouts: Interactions between social context, self-perceptions, school engagement, and student dropout. *Journal of Adolescence, 35,* 787–798. http://dx.doi.org/10.1016/j.adolescence.2011.11.004

Finn, J. D. (1989). Withdrawing from school. *Review of Educational Research, 59,* 117–142. http://dx.doi.org/10.3102/00346543059002117

Finn, J. D., & Zimmer, K. S. (2012). Student engagement: What is it? Why does it matter? In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 97–131). New York, NY: Springer US. http://dx.doi.org/10.1007/978-1-4614-2018-7_5

Furrer, C., & Skinner, E. A. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology, 95,* 148–162. http://dx.doi.org/10.1037/0022-0663.95.1.148

Furrer, C. J., Skinner, E. A., & Pitzer, J. R. (2014). The influence of teacher and peer relationships on students' classroom engagement and everyday resilience. In D. J. Shernoff & J. Bempechat (Eds.), *National Society for the Study of Education yearbook. Engaging youth in schools: Empirically based models to guide future innovations* (Vol. 113, pp. 101–123). New York, NY: Columbia University, Teachers' College.

Goodenow, C. (1993). Classroom belonging among early adolescent students: Relationships to motivation and achievement. *Journal of Early Adolescence, 13,* 21–43. http://dx.doi.org/10.1177/0272431693013001002

Hallinan, M. T., & Williams, R. A. (1990). Students' characteristics and the peer-influence process. *Sociology of Education, 63,* 122–132. http://dx.doi.org/10.2307/2112858

Hamre, B. K., & Pianta, R. C. (2001). Early teacher–child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development, 72,* 625–638. http://dx.doi.org/10.1111/1467-8624.00301

Harris, J. R. (1995). Where is the child's environment? A group socialization theory of development. *Psychological Review, 102,* 458–489. http://dx.doi.org/10.1037/0033-295X.102.3.458

Hughes, J., & Kwok, O. M. (2007). Influence of student–teacher and parent–teacher relationships on lower achieving readers' engagement and achievement in the primary grades. *Journal of Educational Psychology, 99,* 39–51. http://dx.doi.org/10.1037/0022-0663.99.1.39

Kandel, D. B. (1985). On processes of peer influences in adolescent drug use: A developmental perspective. *Advances in Alcohol & Substance Use, 4,* 139–162. http://dx.doi.org/10.1300/J251v04n03_07

Kindermann, T. A. (1993). Natural peer groups as contexts for individual development: The case of children's motivation in school. *Developmental Psychology, 29,* 970–977. http://dx.doi.org/10.1037/0012-1649.29.6.970

Kindermann, T. A. (1996). Strategies for the study of individual development within naturally existing peer groups. *Social Development, 5,* 158–173. http://dx.doi.org/10.1111/j.1467-9507.1996.tb00078.x

Kindermann, T. A. (2003). Development of children's social relationships. In J. Valsiner & K. Connolly (Eds.), *Handbook of developmental psychology* (pp. 407–430). Thousand Oaks, CA: Sage.

Kindermann, T. A. (2007). Effects of naturally existing peer groups on changes in academic engagement in a cohort of sixth graders. *Child Development, 78,* 1186–1203. http://dx.doi.org/10.1111/j.1467-8624.2007.01060.x

Kindermann, T. A. (2011). Commentary: The invisible hand of the teacher. *Journal of Applied Developmental Psychology, 32,* 304–308. http://dx.doi.org/10.1016/j.appdev.2011.04.005

Kindermann, T. A., & Skinner, E. A. (2012). Will the real peer group please stand up? A "tensegrity" approach to examining the synergistic influences of peer groups and friendship networks on academic development. In F. Pajares & T. Urdan (Series Eds.) & A. Ryan & G. Ladd (Vol. Eds.), *Adolescents and education: Peer relationships and adjustment at school* (pp. 51–78). New York, NY: Information Age.

Kiuru, N., Aunola, K., Lerkkanen, M. K., Pakarinen, E., Poskiparta, E., Ahonen, T., . . . Nurmi, J. E. (2015). Positive teacher and peer relations combine to predict primary school students' academic skill development. *Developmental Psychology, 51,* 434–446. http://dx.doi.org/10.1037/a0038911

Klem, A. M., & Connell, J. P. (2004). Relationships matter: Linking teacher support to student engagement and achievement. *Journal of School Health, 74,* 262–273. http://dx.doi.org/10.1111/j.1746-1561.2004.tb08283.x

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. New York, NY: Guilford Press.

Kurdek, L. A., & Sinclair, R. J. (2000). Psychological, family, and peer predictors of academic outcomes in first-through fifth-grade children. *Journal of Educational Psychology, 92,* 449–457. http://dx.doi.org/10.1037/0022-0663.92.3.449

Ladd, G. W. (1990). Having friends, keeping friends, making friends, and being liked by peers in the classroom: Predictors of children's early school adjustment? *Child Development, 61,* 1081–1100. http://dx.doi.org/10.2307/1130877

Lempers, J. D., & Clark-Lempers, D. S. (1992). Young, middle, and late adolescents' comparisons of the functional importance of five significant relationships. *Journal of Youth and Adolescence, 21,* 53–96. http://dx.doi.org/10.1007/BF01536983

Li, Y., & Lerner, R. M. (2011). Trajectories of school engagement during adolescence: Implications for grades, depression, delinquency, and substance use. *Developmental Psychology, 47,* 233–247. http://dx.doi.org/10.1037/a0021307

Libbey, H. P. (2004). Measuring student relationships to school: Attachment, bonding, connectedness, and engagement. *Journal of School Health, 74,* 274–283. http://dx.doi.org/10.1111/j.1746-1561.2004.tb08284.x

Little, T. D., Bovaird, J. A., & Widaman, K. F. (2006). On the merits of orthogonalizing powered and product terms: Implications for modeling interactions among latent variables. *Structural Equation Modeling, 13,* 497–519. http://dx.doi.org/10.1207/s15328007sem1304_1

Luthar, S. S., Cicchetti, D., & Becker, B. (2000). The construct of resilience: A critical evaluation and guidelines for future work. *Child Development, 71,* 543–562. http://dx.doi.org/10.1111/1467-8624.00164

Lynch, A. D., Lerner, R. M., & Leventhal, T. (2013). Adolescent academic achievement and school engagement: An examination of the role of school-wide peer culture. *Journal of Youth and Adolescence, 42,* 6–19. http://dx.doi.org/10.1007/s10964-012-9833-0

Martin, A. J., & Dowson, M. (2009). Interpersonal relationships, motivation, engagement, and achievement: Yields for theory, current issues, and educational practice. *Review of Educational Research, 79,* 327–365. http://dx.doi.org/10.3102/0034654308325583

Moore, S. R., & Depue, R. A. (2016). Neurobehavioral foundation of environmental reactivity. *Psychological Bulletin, 142,* 107–164. http://dx.doi.org/10.1037/bul0000028

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Nelson, R. M., & DeBacker, T. K. (2008). Achievement motivation in adolescents: The role of peer climate and best friends. *Journal of Experimental Education, 76,* 170–189. http://dx.doi.org/10.3200/JEXE.76.2.170-190

Newmann, F. M. (Ed.). (1992). *Student engagement and achievement in American secondary schools*. New York, NY: Teachers College Press.

Nurmi, J. E., & Kiuru, N. (2015). Students' evocative impact on teacher instruction and teacher–child relationships Theoretical background and an overview of previous research. *International Journal of Behavioral Development, 39,* 445–457. http://dx.doi.org/10.1177/0165025415592514

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling, 14,* 535–569. http://dx.doi.org/10.1080/10705510701575396

Osterman, K. F. (2000). Students' need for belonging in the school community. *Review of Educational Research, 70,* 323–367. http://dx.doi.org/10.3102/00346543070003323

Quin, D. (in press). Longitudinal and contextual associations between teacher–student relationships and student engagement: A systematic review. *Review of Educational Research.*

Raufelder, D., Jagenow, D., Drury, K., & Hoferichter, F. (2013). Social relationships and motivation in secondary school: Four different moti-

vation types. *Learning and Individual Differences, 24,* 89–95. http://dx.doi.org/10.1016/j.lindif.2012.12.002

Reeve, J. (2012). A self-determination theory perspective on student engagement. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 149–172). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-2018-7_7

Roeser, R. W., Eccles, J. S., & Sameroff, A. J. (2000). School as a context of early adolescents' academic and social-emotional development: A summary of research findings. *Elementary School Journal, 100,* 443–471. http://dx.doi.org/10.1086/499650

Rumberger, R. W., & Rotermund, S. (2012). The relationship between engagement and high school dropout. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 491–513). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-2018-7_24

Ryan, A. M. (2000). Peer groups as a context for the socialization of adolescents' motivation, engagement, and achievement in school. *Educational Psychologist, 35,* 101–111. http://dx.doi.org/10.1207/S15326985EP3502_4

Ryan, A. M. (2001). The peer group as a context for the development of young adolescent motivation and achievement. *Child Development, 72,* 1135–1150. http://dx.doi.org/10.1111/1467-8624.00338

Ryan, A. M., & Shin, H. (2011). Help-seeking tendencies during early adolescence: An examination of motivational correlates and consequences for achievement. *Learning and Instruction, 21,* 247–256. http://dx.doi.org/10.1016/j.learninstruc.2010.07.003

Ryan, R. M., & Deci, E. L. (2009). Promoting self-determined school engagement: Motivation, learning, and well-being. In K. R. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 171–195). New York, NY: Routledge.

Sabol, T. J., & Pianta, R. C. (2012). Recent trends in research on teacher–child relationships. *Attachment & Human Development, 14,* 213–231. http://dx.doi.org/10.1080/14616734.2012.672262

Sage, N. A., & Kindermann, T. A. (1999). Peer networks, behavior contingencies, and children's engagement in the classroom. *Merrill-Palmer Quarterly, 454,* 143–171.

Shernoff, D. J., Csikszentmihalyi, M., Shneider, B., & Shernoff, E. S. (2003). Student engagement in high school classrooms from the perspective of flow theory. *School Psychology Quarterly, 18,* 158–176. http://dx.doi.org/10.1521/scpq.18.2.158.21860

Skinner, E. A. (2016). Engagement and disaffection as central to processes of motivational resilience and development. In K. R. Wentzel & D. B. Miele (Eds.), *Handbook of motivation at school* (pp. 145–168). New York, NY: Routledge.

Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85,* 571–581. http://dx.doi.org/10.1037/0022-0663.85.4.571

Skinner, E. A., Kindermann, T. A., Connell, J. P., & Wellborn, J. G. (2009). Engagement and disaffection as organizational constructs in the dynamics of motivational development. In K. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 223–245). London, UK: Routledge.

Skinner, E. A., Kindermann, T. A., & Furrer, C. (2009). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and Psychological Measurement, 69,* 493–525. http://dx.doi.org/10.1177/0013164408323233

Skinner, E. A., Wellborn, J. G., & Connell, J. P. (1990). What it takes to do well in school and whether I've got it: A process model of perceived control and children's engagement and achievement in school. *Journal of Educational Psychology, 82,* 22–32. http://dx.doi.org/10.1037/0022-0663.82.1.22

Skinner, E. A., Zimmer-Gembeck, M. J., Connell, J. P., Eccles, J. S., & Wellborn, J. G. (1998). Individual differences and the development of perceived control. *Monographs of the Society for Research in Child Development, 63,* i–vi, 1–220. http://dx.doi.org/10.2307/1166220

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). New York, NY: Allyn & Bacon.

Ullah, H., & Wilson, M. A. (2007). Students' academic success and its association to student involvement with learning and relationships with faculty and peers. *College Student Journal, 41,* 1192–1202.

Wang, M. T., & Degol, J. (2014). Staying engaged: Knowledge and research needs in student engagement. *Child Development Perspectives, 8,* 137–143. http://dx.doi.org/10.1111/cdep.12073

Wang, M. T., & Eccles, J. S. (2012). Social support matters: Longitudinal effects of social support on three dimensions of school engagement from middle to high school. *Child Development, 83,* 877–895. http://dx.doi.org/10.1111/j.1467-8624.2012.01745.x

Wang, M. T., & Fredricks, J. A. (2014). The reciprocal links between school engagement, youth problem behaviors, and school dropout during adolescence. *Child Development, 85,* 722–737. http://dx.doi.org/10.1111/cdev.12138

Wang, M. T., & Peck, S. C. (2013). Adolescent educational success and mental health vary across school engagement profiles. *Developmental Psychology, 49,* 1266–1276. http://dx.doi.org/10.1037/a0030028

Weiner, B. (1990). History of motivational research in education. *Journal of Educational Psychology, 82,* 616–622. http://dx.doi.org/10.1037/0022-0663.82.4.616

Wellborn, J. (1992). *Engaged and disaffected action: The conceptualization and measurement of motivation in the academic domain.* Unpublished doctoral dissertation, Department of Psychology, University of Rochester, Rochester, NY.

Wentzel, K. R. (1994). Relations of social goal pursuit to social acceptance, classroom behavior, and perceived social support. *Journal of Educational Psychology, 86,* 173–182. http://dx.doi.org/10.1037/0022-0663.86.2.173

Wentzel, K. R. (1997). Student motivation in middle school: The role of perceived pedagogical caring. *Journal of Educational Psychology, 89,* 411–419. http://dx.doi.org/10.1037/0022-0663.89.3.411

Wentzel, K. R. (1999). Social–motivational processes and interpersonal relationships: Implications for understanding motivation at school. *Journal of Educational Psychology, 91,* 76–97. http://dx.doi.org/10.1037/0022-0663.91.1.76

Wentzel, K. R. (2009a). Peers and academic functioning at school. In K. Rubin, W. Bukowski, & B. Laursen (Eds.), *Handbook of peer interactions, relationships, and groups. Social, emotional, and personality development in context* (pp. 531–547). New York, NY: Guilford Press.

Wentzel, K. R. (2009b). Students' relationships with teachers as motivational contexts. In K. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 301–322). London, UK: Routledge.

Wentzel, K. R., Battle, A., Russell, S. L., & Looney, L. B. (2010). Social supports from teachers and peers as predictors of academic and social motivation. *Contemporary Educational Psychology, 35,* 193–202. http://dx.doi.org/10.1016/j.cedpsych.2010.03.002

Wentzel, K. R., & Miele, D. (Eds.). (2016). *Handbook of motivation at school* (2nd ed.). Mahwah, NJ: Erlbaum.

Wentzel, K. R., & Muenks, K. (2016). Peer influence on students' motivation, academic achievement, and social behavior. In K. Wentzel & G. Ramani (Eds.), *Handbook of social influences in school contexts: Social–emotional, motivation, and cognitive outcomes* (pp. 13–30). New York, NY: Routledge.

Wentzel, K. R., & Ramani, G. B. (Eds.). (2016). *Handbook of social influences in school contexts: Social–emotional, motivation, and cognitive outcomes.* New York, NY: Routledge.

Wentzel, K. R., & Watkins, D. E. (2011). Instruction based on peer

interactions. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 322–343). New York, NY: Routledge.

Wigfield, A., Eccles, J. S., Fredricks, J. A., Simpkins, S., Roeser, R., & Schiefele, U. (2015). Development of achievement motivation and engagement. In R. M. Lerner (Series Ed.) & M. Lamb (Vol. Ed.), *Handbook of child psychology and developmental science: Vol. 3. Socioemotional processes* (7th ed., pp. 657–700). New York, NY: Wiley. http://dx.doi.org/10.1002/9781118963418.childpsy316

Received January 22, 2015
Revision received November 9, 2016
Accepted November 11, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 5, 653–665

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000161

It's All a Matter of Perspective: Viewing First-Person Video Modeling Examples Promotes Learning of an Assembly Task

Logan Fiorella
University of Georgia

Tamara van Gog and Vincent Hoogerheide
Utrecht University and Erasmus University Rotterdam

Richard E. Mayer
University of California Santa Barbara

The present study tests whether presenting video modeling examples from the learner's (first-person) perspective promotes learning of an assembly task, compared to presenting video examples from a third-person perspective. Across 2 experiments conducted in different labs, university students viewed a video showing how to assemble an 8-component circuit on a circuit board. Students who viewed the assembly video recorded from a first-person perspective performed significantly better than those who viewed the video from a third-person perspective on accuracy in assembling the circuit in both experiments and on time to assemble the circuit in Experiment 1, but not in Experiment 2. Concerning boundary conditions, the perspective effect was stronger for more complex tasks (Experiment 1), but was not moderated by imitating the actions during learning (Experiment 1) or explaining how to build the circuit during the test (Experiment 2). This work suggests a perspective principle for instructional video in which students learn better when video reflects a first-person perspective. An explanation based on embodied theories of learning and instruction is provided.

Keywords: video, modeling examples, multimedia learning, perspective taking, embodied cognition

Consider an instructional video showing how to perform a manual task, such as how to construct a circuit on a circuit board. The main goal of this study is to examine techniques for improving the effectiveness of instructional videos, particularly the role of the perspective from which the video is recorded (i.e., first-person or third-person). In two experiments, we examine whether students learn better from an instructional video recorded from a first-person perspective, and whether there are boundary conditions for any perspective effects.

There is rapidly growing interest in the use of video modeling examples for instruction within formal (e.g., online courses) and informal (e.g., YouTube) educational settings, likely due to their convenience, relatively low cost, and high accessibility. A video modeling example involves a human model demonstrating and/or explaining to a learner how to perform a task (van Gog & Rummel, 2010). For example, a student taking an online statistics course may watch videos of an instructor solving problems on a whiteboard, or a person may watch a YouTube video of someone modeling how to tie a necktie or how to play a musical instrument. However, despite their wide implementation, there is relatively little systematic research investigating how to effectively design video lessons.

Observational Learning From Video Modeling Examples

Much of the existing research on learning from modeling examples concerns the effects of different characteristics of the human models (or animated agents) on learning. For example, in a classic study, Schunk, Hanson, and Cox (1987) manipulated the gender of the model and whether the model used an automatic mastery strategy or a more effortful coping strategy to solve math problems. More recent research has further explored the effects of model characteristics, including the model's gender (Hoogerheide, Loyens, & van Gog, 2016) and the model's age and expertise (Hoogerheide, van Wermeskerken, Loyens, & van Gog, 2016). In addition, design issues for instructional video that have been addressed recently include the visibility of the model's face (Kizilcec, Bailenson, & Gomez, 2015; van Gog, Verveer, & Verveer, 2014), the availability of gaze and gesture cues provided by the model (Ouwehand, van Gog, & Paas, 2015), the visibility of the model's hands in a motor task (Castro-Alonso, Ayres, & Paas, 2015; Marcus, Cleary, Wong, & Ayres, 2013), and whether the

This article was published Online First November 10, 2016.

Logan Fiorella, Department of Educational Psychology, University of Georgia; Tamara van Gog and Vincent Hoogerheide, Department of Education, Utrecht University, and Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam; Richard E. Mayer, Department of Psychological and Brain Sciences, University of California Santa Barbara.

Tamara van Gog was supported by a Vidi grant from the Netherlands Organization for Scientific Research (NWO 452-11-06). Richard E. Mayer was supported by grants from the Office of Naval Research (N000141110225 and N0001416112046). We thank Arie Koster, Amanda Seaborne, and Kayla Shires for their assistance with Experiment 1, and Susan Ravensbergen, Kirsten Versijde, and Charly Eielts for their assistance with Experiment 2.

Correspondence concerning this article should be addressed to Logan Fiorella, Department of Educational Psychology, University of Georgia, 110 Carlton Street, Athens, GA 30602. E-mail: lfiorella@uga.edu

model physically draws out diagrams by hand during a lesson (Fiorella & Mayer, 2016). In short, past research has focused primarily on the effects of manipulating the appearance of the model and the model's actions provided in modeling examples.

The current study focuses on a largely ignored but pervasive design feature of video modeling examples: the perspective from which the video is recorded. Although perspective is typically not an issue in lecture-style modeling examples, in which the model is standing next to a screen on which slides are projected illustrating each step in the task (cf. Fiorella & Mayer, 2016; Hoogerheide, van Wermeskerken, et al., 2016; Ouwehand et al., 2015), it may play a role in demonstrations in which objects are being manipulated (cf. Castro-Alonso et al., 2015; Marcus et al., 2013; van Gog et al., 2014). Thus, perspective is a potentially important design consideration for instruction involving concrete manipulatives, commonly used to teach math and science concepts (e.g., Marley & Carbonneau, 2015).

In the current study, we tested whether students would benefit more from observing instructional videos from first-person perspective—that is, with the model performing the task from the perspective of the person observing the task—than from a third-person perspective. We also tested whether the potential effects of perspective would depend on the complexity of the to-be-learned task, and whether engaging in common and effective learning strategies—imitating in Experiment 1 and explaining in Experiment 2—would compensate for the expected detrimental effects of a third-person perspective on performance. Examining potential interactions among task complexity, learning strategies, and instructional methods is valuable because it provides insight into the robustness and generalizability of the findings. Although there is little research investigating the effects of perspective in educational videos (e.g., Lindgren, 2012), basic research in cognitive science supports the proposal that processing material from a first-person perspective may provide important cognitive benefits.

Perspective and Observational Learning

Observing the actions of others can be a powerful way to learn, likely because of the evolutionary benefits of observing and (when the outcome is desirable) imitating other people's actions (Bandura, 1977, 1986; Paas & Sweller, 2012; Sweller & Sweller, 2006). In observational learning, learners must actively interpret the actions of a human model by constructing a cognitive representation of the modeled behavior that is integrated with their prior knowledge (Bandura, 1986). Some have further proposed that this process is facilitated via activation of the mirror neuron system, which generally involves the idea that brain areas activated when performing actions are also activated when observing others perform those actions (Rizzolatti & Craighero, 2004; van Gog, Paas, Marcus, Ayres, & Sweller, 2009). When observing to-be-performed actions from the third-person perspective, learners must mentally transform the representation into their own perspective, such as by translating the model's view of left to their own left. Although humans have the unique ability to take the spatial perspective of others, such mental transformations can be cognitively demanding (Hegarty & Waller, 2004; Kessler & Thomson, 2010). This extraneous load on working memory may be reduced when the model demonstrates the task from the observer's own point of view.

Basic research in cognitive science supports a facilitative effect for processing visuomotor information from the first-person (compared to third-person) perspective. In a study by Vogt, Taylor, and Hopkins (2003), participants were asked to perform a simple hand action after being primed with pictures of hands performing either congruent or incongruent actions presented from the first- or third-person perspective. When participants were provided with a preview of the hand's start position before viewing the prime, only participants who viewed the primes from the first-person perspective were faster at performing the action when the prime displayed a congruent action compared to an incongruent action. The authors concluded that viewing body parts presented in the first-person perspective activates motor planning processes in the observer, which enhances the processing of the visual information associated with the prepared actions.

Kelly and Wheaton (2013) found further support for the notion that first-person perspective enhances motor planning and judgment. Participants were shown images of hands performing movements with tools from either a first-person or third-person perspective, and they were asked to judge the outcome of the action. Action judgments were fastest and most accurate when stimuli were viewed from the first-person perspective, again suggesting that actions are better represented when viewed from the observer's own perspective.

Next to motor planning, there is evidence from research using functional MRI (fMRI) that participants are prepared for later imitation because observing actions activates the motor neurons they would use when performing the actions (Jackson, Meltzoff, & Decety, 2006). Participants viewed video clips of simple hand and foot actions presented from the first- or third-person perspective. Some participants watched the videos passively and others imitated the actions. Behavioral data indicated that response latency to imitate the actions was shorter for the first-person perspective. Further, fMRI data indicated more activity in the left sensory-motor cortex (which would be active when executing the movement oneself) for the first-person perspective compared to third-person perspective, even when participants passively observed and did not imitate the actions. These data are consistent with embodied views of cognition (Barsalou, 2008; Wilson, 2002), which posit that human perception, cognition, and action are closely linked and grounded in one's interactions with the physical world. That is, the sensory-motor system appears more involved in processing actions from the first-person perspective, whereas the third-person perspective requires visuospatial transformations that consume limited processing capacity.

Further evidence that such transformations take time (and may result in errors) comes from a study involving visual perspective taking by Kockler and colleagues (2010). Participants were asked to make judgments about the spatial location of a static or dynamic object from their own perspective (first-person) or from the perspective of a virtual character (third-person). Results indicated that judgments were faster and more accurate when participants were asked to report the location from their own perspective. Further, fMRI data indicated that judgments of the dynamic objects from the first-person perspective resulted in increased activation in the intraparietal sulcus (IPS), an area involved in action preparation. Thus, viewing dynamic stimuli from the first-person perspective appears to improve performance by inducing a readiness to act.

Many other basic behavioral and neuroscience studies support a beneficial effect of viewing, imitating, and judging actions observed from the first-person perspective (e.g., Lorey et al., 2009; Maeda, Kleiner-Fisman, & Pascual-Leone, 2002; Surtees & Apperly, 2012; Vogeley & Fink, 2003). Similarly, the spatial cognition and navigation literatures demonstrate the high cognitive demands associated with spatial perspective taking, showing that that performance typically decreases as the angular disparity between the first-person and target viewpoint increases (e.g., Kozhevnikov, Motes, Rasch, & Blajenkova, 2006; Richardson, Montello, & Hegarty, 1999).

Unfortunately, however, research on the consequences of those findings from basic cognitive science for education and training is scarce. That is, prior research has mainly looked at effects of perspective on performance, not on learning (i.e., later performance in the absence of the observed stimuli; for an exception on spatial learning, see Richardson et al., 1999). Although prior research in educational psychology has involved video lessons presented from a first-person perspective (e.g., Ayres, Marcus, Chan, & Qian, 2009) or a third-person perspective (e.g., Arguel & Jamet, 2009), these studies did not focus on comparing the effects of video lessons presented via different perspectives. One exception is an experiment by Lindgren (2012), in which students interacted within a virtual safety training simulation from either first-person perspective or the perspective of a virtual character (i.e., third-person). Results indicated that participants who received the first-person perspective training performed better on a diagramming task, had better memory for the tasks of the simulation, committed fewer errors, and showed less help-seeking than participants who received the third-person perspective training. Lindgren concluded that virtual environments provide a unique ability to help students adopt a more embodied learning stance, allowing students to interact with learning material from their own point of view.

Overall, the available research evidence suggests a facilitation effect for processing dynamic visual information from the first-person (as opposed to third-person) perspective—consistent with the claim of embodied theories of cognition (Barsalou, 2008; Wilson, 2002) that the first-person perspective uniquely serves to shape one's cognitive representations of space and action. Accordingly, viewing materials from a third-person perspective requires learners to generate additional visuospatial transformations in order to translate observed actions into their own perspective, which creates extraneous cognitive load that consequently impairs performance. Open questions remain regarding the applicability of this basic finding to educational settings in which the focus is on learning outcomes, including potential boundary conditions associated with features of the to-be-learned task and actions of the student during learning.

The Present Study

The main aim of the current study was to investigate the hypothesis derived from the literature reviewed above, that presenting video modeling examples of an assembly task from the performer's (first-person) perspective would result in better learning (as assessed by speed and accuracy of subsequent assembly performance) than presenting videos from the third-person perspective. We conducted two experiments (in two different labs), in which university students viewed narrated video examples showing a model's hands performing an assembly task involving electric circuits. Half of the students viewed videos presented from the third-person perspective (third-person group), whereas the other half viewed videos presented from the first-person perspective (first-person group). Then, all students assembled the circuits on their own (from their perspective). According to the hypothesis, a main effect of perspective was expected in both experiments, with the first-person perspective outperforming the third-person perspective (i.e., faster and more accurate assembly).

A second aim of this study was to further explore the conditions under which video perspective influences subsequent assembly performance. Experiment 1 tested whether the effects of perspective are moderated by task complexity (within-subjects) and whether learners imitated the model during learning (between-subjects). Using only the complex tasks, Experiment 2 tested whether the effects of perspective were moderated by whether learners gave a verbal explanation while they assembled the circuit (between-subjects).

With regard to task complexity (Experiment 1), it was expected that the hypothesized beneficial effects of the first-person perspective would show primarily on complex tasks. That is, on simple tasks, which involve fewer interacting elements, overall working memory load is lower (Sweller, Ayres, & Kalyuga, 2011), and any additional processing demands imposed by the third-person perspective could be accommodated without hampering learning. The expected interaction between perspective and task complexity should also correspond to students' subjective ratings of mental effort during learning, with the highest levels of mental effort occurring when students view high-complexity tasks from the third-person perspective.

As for imitation (in Experiment 1), it was hypothesized that imitating the steps during example study might reduce reassembly time and effort, and boost test performance, for both perspectives compared to no imitation (i.e., main effect of imitation), because it would lead to deeper example processing and allow learners to practice during example study. Whereas fundamental research shows that imitation might be easier when seeing a first-person than a third-person view (Watanabe, Higuchi, & Kikuchi, 2013), it was expected that imitation might compensate for the expected negative effects of the third-person perspective (i.e., interaction effect of perspective and imitation). Although imitation is not necessary for observational learning to occur, it can aid in the process of converting symbolic codes acquired through observation into appropriate actions (Bandura, 1986). Having performed the actions (from the first-person perspective) allows for consolidating the first-person action in memory instead of the observed third-person action. Moreover, performing the actions oneself during learning may aid in transforming the observed third-person actions into first-person action representations. Without imitation, this transformation has to be made mentally. Performing the action during imitation, however, allows the learner to partially offload that mental transformation onto the external environment (e.g., by rotating the objects), thereby reducing working memory demands (Kirsh & Maglio, 1994). So, if assembly test performance is slower and less accurate in the third-person perspective condition than in the first-person perspective condition, then having made that translation during learning would boost their test performance compared to the no imitation third-person condition.

A similar expectation applied to explaining (in Experiment 2). Based on research on learning by explaining, which indicates that generating explanations is an effective learning strategy (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Fiorella & Mayer, 2015a, 2015b), it was expected that instructing participants that they would have to explain how to build the circuit afterward, might boost their performance compared to no explaining instruction under both perspectives (i.e., main effect of explaining). That is, knowing that they would have to give the explanation themselves later on, might result in deeper processing of the example, and especially the model's verbal explanation. Moreover, this "imitation" of the verbal explanation by the model (which was always from the first-person perspective)—by giving the same explanation to another (fictitious, nonpresent) student while assembling the circuit—might guide their assembly test performance and help compensate for detrimental effects on test performance in the third-person perspective condition (i.e., interaction effect between perspective and explaining). That is, similarly to imitating during learning, explaining during reassembly might help learners better align the model's actions and verbal instructions with their own perspective, particularly when the model's actions are presented from the third-person perspective. Taken together, the two experiments address whether engaging in learning strategies alleviates the increased processing demands expected from viewing instructional videos from the instructor's perspective.

Experiment 1

Method

Participants and design. The participants were 105 university students from the Psychology Subject Pool of a university in the United States who participated to fulfill a course requirement. The mean age of participants was 19.30 years (SD = 1.32), and there were 73 women and 32 men. Participants were randomly assigned to one of four conditions, based on two between-subjects factors—perspective of the instructional videos (first-person or third-person) and whether or not students imitated the video model's actions during learning (imitate or no-imitate). There were 26 students in the first-person/imitate group, 26 in the first-person/no-imitate group, 25 in the third-person/imitate group, and 28 in the third-person/no-imitate group. The groups did not significantly differ in terms of average age, number of women/men, handedness, or prior experience (as indicated by a self-report checklist described below). Task complexity (low or high) served as a within-subjects factor and was counterbalanced across conditions.

Materials. The paper-based materials consisted of a consent form, a demographics questionnaire, and a mental effort rating scale.[1] The consent form described the details of the study, informed participants that they would be videotaped during the experiment and that their privacy was protected, and included a place for them to sign. The demographics form asked participants to provide their age, gender, and handedness. Students also rated their relevant prior experience by placing a check mark next to each of eight items that apply to them, such as "I have taken a college-level course in physics," "I have worked on a circuit board," "I have installed a new light switch or electrical outlet," and "I know the difference between serial and parallel circuits."

The mental effort rating scale (Paas, 1992) asked participants to rate how much mental effort they invested while completing a particular task (e.g., watching an instructional video, building an electric circuit). Students recorded their response on a 9-point scale ranging from *Extremely low mental effort* to *Extremely high mental effort*. This common form of assessing mental effort has been shown to be sensitive enough to detect objective variations in task complexity (Ayres, 2006; Paas & Van Merrienboer, 1994).

The learning task materials consisted of a model electrical circuit kit—called Electronic Snap Circuits (by Elenco)—designed to teach students about how electrical circuits work. Students learn how to build electrical circuits by connecting (i.e., "snapping") different components (e.g., batteries, resistors, wires, LED lights) to a circuit board and to each other. In the current study, students learned how to build two circuit configurations—a low complexity circuit (shown in Figure 1) and a high complexity circuit (shown in Figure 2). As shown in the figure, both circuit configurations contain a total of eight components. However, the high complexity circuit contains more unique components (6) than the low complexity circuit (5), and the high complexity circuit contains components that must be placed in a specific orientation in order for the circuit to work. For example, in the high complexity circuit, the red LED light must point toward the green LED light, and the green LED light must point toward the battery. There are no such orientation requirements for the components in the low complexity circuit.

There were two computer-based instructional videos—a first-person version (exemplified in Figure 3) and a third-person version (exemplified in Figure 4). The instructional videos showed a male model's hands demonstrating how to build the low-complexity and high-complexity circuits while he provided narrated instructions for each of the eight steps.

The first-person perspective video showed the model's hands, as they would appear if the observer of the video were completing the task. The third-person perspective video showed the model's hands, as they would appear if someone facing the observer were completing the task. As the model placed each of the eight components on the circuit board, the oral instructions identified a component and described where it should be placed in relation to other components on the board—for example: "Place the switch below the right end of the long wire . . ." The videos were segmented to pause after the model completed a step. Students who were assigned to one of the imitating conditions would then imitate the step using their own model circuit kit before clicking to continue the video to the next step, whereas students who were assigned to one of the none-imitating conditions would simply click to continue the video to the next step. The videos were recorded simultaneously from the first-person perspective and from the third-person perspective to create identical versions from both perspectives. The low-complexity video lasted 82 seconds (excluding pauses) and contained 94 spoken words, whereas the high-complexity video lasted 90 seconds (excluding pauses) and contained 160 spoken words.

[1] We also asked participants to complete a perspective-taking test (Hegarty & Waller, 2004) upon completion of the demographics questionnaire; however, it did not significantly correlate with any of the dependent measures, and so it was not included in the analyses.

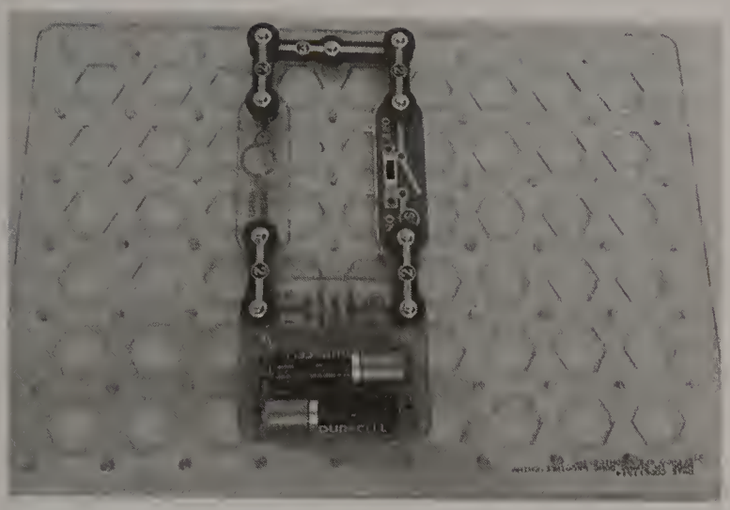


Figure 1. Low-complexity circuit. See the online article for the color version of this figure.

Participants were assessed on their ability to assemble the low- and high-complexity circuits on their own after watching the respective instructional video. During assembly, participants were provided with the eight components needed to build the circuit along with five distractor components. Performance measures consisted of total time to assemble the circuit, accuracy at rebuilding the circuit, and frequency of three types of assembly errors. Participants were asked to assemble the circuit exactly as they saw in the video and were informed that they would be timed. Assembly time was measured from the time participants started to assemble the circuit until they stated they were finished or could not complete any more. Assembly accuracy was measured by totaling the number of correct circuit components in the correct locations and orientations on the circuit board, out of a possible 8 points for each circuit. For Experiment 1, two raters scored participants' assembly accuracy blind to experimental conditions, yielding high interrater reliability (low-complexity circuit: $r = .85$; high-complexity circuit: $r = .82$). Any discrepancies between raters were settled by consensus. For Experiment 2, there was 100% agreement between two raters based on 10% of the data, and so one rater scored the remaining data.

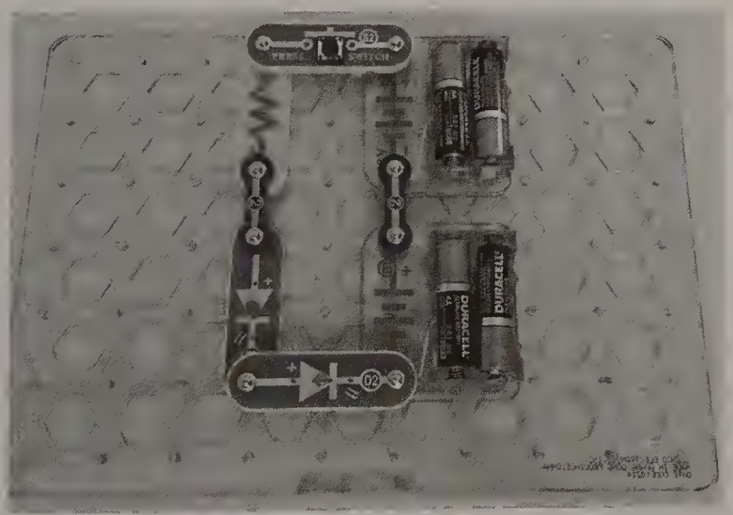


Figure 2. High complexity circuit. See the online article for the color version of this figure.

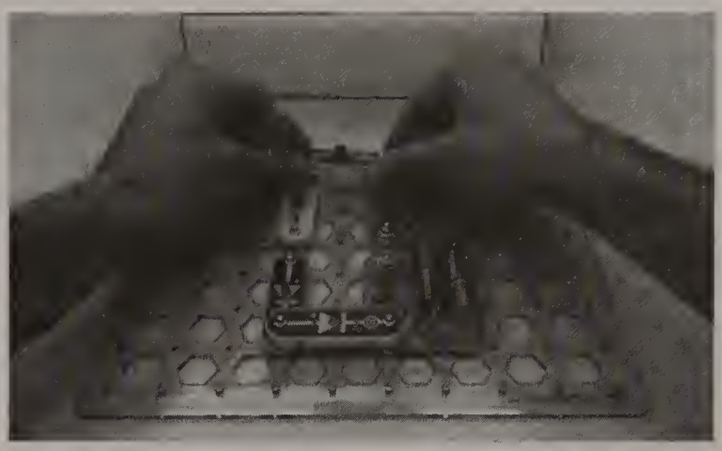


Figure 3. Screenshot from first-person instructional video (high-complexity task). See the online article for the color version of this figure.

Assembly errors were measured by coding whether a participant committed three types of errors in their assembly of the circuit: perspective errors, location errors, or component errors. A perspective error consisted of assembling the circuit from the incorrect perspective (i.e., third-person). A location error consisted of assembling the circuit on the incorrect location on the circuit board. Finally, a component error consisted of using a component that does not make up that circuit (e.g., using a flip switch instead of a press switch). There was 100% agreement between two raters based on 10% of the data for Experiment 1, and 97% agreement between two raters based on 10% of the data for Experiment 2.

The recognition test was a lab-developed computer-based test intended to provide an additional assessment following assembly of the high-complexity circuit. The test presented participants with a series of 40 photos (one at a time) of correct or incorrect versions of the high-complexity circuit at four different orientations (i.e., first-person, 90 degrees turned left, third-person, or 90 degrees turned right). There was a delay of approximately 500 ms between each of the trials. Participants were required to determine whether each photo was the same or different from the circuit that they had just reassembled. Half of the trials were "same" trials; the other half were "different" trials, in which the photos were of a circuit

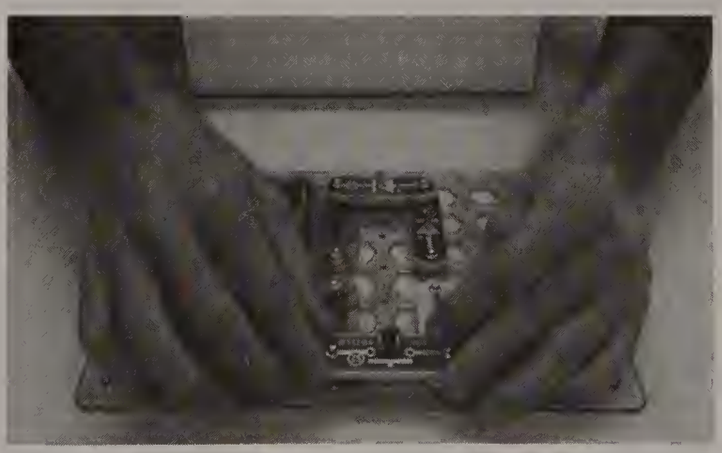


Figure 4. Screenshot from third-person instructional video (high-complexity task). See the online article for the color version of this figure.

with one component in the incorrect location (e.g., location of the resistor and red LED light switched). All trials were distributed evenly across the four orientations and presented randomly to participants using direct reaction time (RT). Performance was assessed via accuracy (out of 40) and average response time. Recognition test accuracy (but not response time) significantly correlated with assembly time, $r = -.33$, $p < .001$ and accuracy, $r = .30$, $p = .002$ for Experiment 1, providing evidence for the validity of the measure.

Apparatus. The apparatus consisted of two Dell computers with 17-in. screens, two Cyber Acoustics headphones, and two web cameras.

Procedure. The participants were randomly assigned to conditions, and were tested up to two per session in individual lab cubicles. After they provided informed consent, they completed the demographics questionnaire. Then, participants watched the first instructional video demonstrating how to build either the low- or high-complexity circuit (for counterbalancing), from either the first-person or third-person perspective (for the perspective variable). Those assigned to imitate conditions completed each step along with the video model using the electric circuit kit; those assigned to no-imitate conditions watched the video without imitating the video model (for the imitate variable). After watching the instructional video, participants completed the mental effort rating scale and then were asked to assemble the circuit on their own using the electric circuit kit. After attempting to assemble the circuit, participants again completed the mental effort rating scale. The same procedure was repeated for the second instructional video (i.e., low- or high-complexity, based on counterbalancing). The order of instructional videos was counterbalanced across conditions. After participants assembled the high-complexity circuit, participants completed the recognition test, followed by the mental effort rating scale. The total duration of the experiment was approximately 60 min.

Results and Discussion

Due to a technical issue with video recording, we do not have data for one participant's accuracy performance on the low complexity circuit. This participant's data is excluded from the relevant analyses presented below. Partial eta squared is reported as a measure of effect size, with values of .01, .06, and .14 generally representing a small, medium, and large effect size, respectively (Cohen, 1988).

Do students learn better from videos recorded from a first-person perspective than from a third-person perspective? The primary research question addressed in this study concerns whether students learn better from instructional videos presented in a first-person perspective than in a third-person perspective. Table 1 shows the mean and standard deviation for the four groups on number of correctly placed components (accuracy) on the low- and high-complexity assembly tasks. A mixed factorial analysis of variance (ANOVA) was conducted, with perspective (first-person or third-person), imitation (imitate or no-imitate), and circuit order (low-high complexity or high-low complexity) serving as between-subjects factors, and circuit complexity (low or high) serving as a within-subjects factor, and number of correctly placed components (out of 8) on the assembly tasks as the dependent measure. Consistent with predictions, there was a significant main effect of perspective, $F(1, 96) = 6.38$, $p = .013$, $\eta_p^2 = .06$, in which students correctly placed more components on the assembly tasks after viewing a first-person video ($M = 7.59$, $SD = 0.86$) than a third-person video ($M = 7.18$, $SD = 0.87$).

Table 1 also shows the mean and standard deviation for the four groups on number of seconds taken for the low- and high-complexity assembly tasks. A mixed factorial analysis of variance (ANOVA) was conducted, with perspective (first-person or third-person), imitation (imitate or no-imitate), and circuit order (low-high complexity first or high-low complexity first) serving as between-subjects factors, and circuit complexity (low or high) serving as a within-subjects factor, and total assembly time serving as the dependent measure. Consistent with predictions, there was a significant main effect of perspective, $F(1, 97) = 6.34$, $p = .013$, $\eta_p^2 = .06$, in which students completed the assembly tasks faster after viewing a first-person video ($M = 68.91$, $SD = 47.02$) than a third-person video ($M = 92.06$, $SD = 47.18$).

Table 1
Means (and SD) of Assembly Time and Accuracy per Condition for Experiment 2

	First-person perspective		Third-person perspective	
	No imitation	Imitation	No imitation	Imitation
Assembly accuracy				
Low complexity	7.73 (.67)	7.84 (.37)	7.68 (.86)	7.72 (.68)
High complexity	7.19 (1.10)	7.60 (.91)	6.57 (1.97)	6.76 (1.59)
Assembly time (s)				
Low complexity	67.38 (58.45)	45.23 (10.78)	67.61 (39.04)	54.80 (18.71)
High complexity	99.77 (99.15)	63.27 (24.57)	131.93 (95.86)	113.04 (90.81)
Recognition test				
Time	4490.66 (1,464.12)	3658.82 (1,051.98)	4195.03 (1,509.02)	4088.87 (1,109.96)
Accuracy	90.19 (14.90)	90.00 (12.37)	87.41 (15.10)	84.20 (16.55)
Mental effort (1–9)				
Example study (low complexity)	5.69 (1.78)	4.35 (1.52)	6.46 (1.35)	3.96 (1.88)
Assembly test (low complexity)	5.00 (1.92)	3.62 (1.60)	5.57 (1.67)	4.04 (2.01)
Example study (high complexity)	6.04 (1.31)	4.77 (1.82)	6.64 (1.25)	5.24 (2.26)
Assembly test (high complexity)	5.50 (1.84)	4.85 (2.22)	6.14 (1.53)	5.40 (2.16)
Recognition test	5.62 (1.42)	5.12 (1.56)	6.14 (1.51)	5.56 (1.92)

Overall, these results provide support for a perspective effect: People learn better from instructional videos recorded from a first-person perspective than from a third-person perspective. This is the primary finding of Experiment 1.

Does the perspective effect depend on the complexity of the task? A secondary question concerns whether the perspective effect favoring first-person videos is stronger for more complex assembly tasks. The ANOVA on the number of correctly placed components on the test (as summarized in Table 1) yielded a significant perspective by complexity interaction, $F(1, 96) = 4.57$, $p = .035$, $\eta_p^2 = .05$, in which the perspective effect was present in the high-complexity task (first-person video group: $M = 7.39$, $SD = 1.02$; third-person video group: $M = 6.66$, $SD = 1.79$) but not in the low-complexity task (first-person video group: $M = 7.78$, $SD = 0.54$; third-person video group: $M = 7.70$; $SD = 0.77$). Similarly, the ANOVA on the total assembly time (as summarized in Table 2) yielded a significant perspective-by-complexity interaction, $F(1, 97) = 4.79$, $p = .031$, $\eta_p^2 = .05$, in which the perspective effect was stronger for the high-complexity task (first-person video group: $M = 81.52$, $SD = 73.85$; third-person video group $M = 123.02$, $SD = 93.10$) than for the low-complexity task (first-person video group: $M = 56.31$, $SD = 43.09$; third-person group: $M = 61.57$, $SD = 31.54$). Overall, these data are consistent with the second hypothesis that the perspective effect is strong for high-complexity tasks but not for low-complexity tasks. Thus, task complexity appears to be a potential boundary condition (or moderator) for the perspective effect.

As expected, performance accuracy was significantly better on low-complexity tasks ($M = 7.74$, $SD = 0.71$) than on high-complexity tasks ($M = 7.03$, $SD = 1.43$), $F(1, 96) = 21.72$, $p < .001$, $\eta_p^2 = .19$, and assembly time was significantly shorter on low-complexity tasks ($M = 58.77$, $SD = 35.56$) than on high-complexity tasks ($M = 102.20$, $SD = 82.49$), $F(1, 97) = 27.22$, $p < .001$, $\eta_p^2 = .22$.

Does the perspective effect depend on whether students imitated the video during learning? A third question concerns whether the perspective effect favoring first-person videos is stronger when students do not have the opportunity to imitate the instructor's steps on assembly tasks. The ANOVAs showed no significant interaction between imitating and perspective for assembly accuracy $F(1, 96) < 1$, $p = .637$, or assembly time, $F(1, 97) < 1$, $p = .451$, indicating no support for the idea that imitating might compensate for the negative effects of a third-person perspective. Overall, there is no evidence for the third hypothesis that imitating during learning serves as a boundary condition (or moderator) for the perspective effect.

Imitation did, however, yield a significant main effect in which students who imitated during learning ($M = 69.30$, $SD = 47.06$) performed better on assembly time than students who not imitate ($M = 91.67$, $SD = 48.43$), $F(1, 97) = 5.93$, $p = .017$, $\eta_p^2 = .06$; however, there was not a significant main effect of imitation for assembly accuracy, $F(1, 96) = 1.30$, $p = .258$, $\eta_p^2 = .01$.

Does perspective affect the type of errors students make on assembly tasks? As a follow-up to the first research question, we also analyzed the frequency at which students made three different types of errors during assembly: perspective errors, location errors, and component errors. Perspective errors involve reassembling the circuit from the third-person perspective rather than from the first-person perspective; location errors involve reassembling the circuit on the incorrect location on the circuit board grid coordinates; and component errors involve reassembling the circuit using components that do not make up that circuit (e.g., using a flip switch instead of a press switch).

Two-sided chi-square tests were conducted to analyze the number of each type of error across perspective (first-person or third-person) and circuit complexity (low or high). For the low-complexity task, students who viewed the videos from the third-person perspective (8 out of 53, or 15.1%) were significantly more likely to commit perspective errors than students who viewed the videos from the first-person perspective (0/51, or 0%), $\chi^2(1) = 8.34$, $p = .004$. The third-person perspective group (11/53, or 20.1%) was also significantly more likely to make location errors than the first-person group, (2/51, or 3.9%), $\chi^2(1) = 6.73$, $p = .009$. The groups did not significantly differ in number of component errors (first-person: 8/51, or 15.7%; third-person: 6/53, or 11.3%; $\chi^2(1) = 0.43$, $p = .514$).

The same pattern of data was found for the high-complexity task. The third-person perspective group made significantly more perspective errors (13/53, or 24.5%) and location errors (13/53, or 24.5%) compared to the first-person group (perspective: 0/52; $\chi^2(1) = .003$, $p = .955$; location: 3/52, or 5.8%; $\chi^2(1) = 7.15$, $p = .007$), and the groups did not significantly differ on number of component errors (first-person: 13/52, or 25.0%; third-person: 13/53, or 24.5%; $\chi^2(1) = 0.003$, $p = .955$). Overall, viewing

Table 2
Means (and SD) of Assembly Time, Accuracy, Effort Ratings, and Error Types per Condition for Experiment 1

	First-person perspective		Third-person perspective	
	No explanation	Explanation	No explanation	Explanation
Assembly test				
Accuracy	6.65 (1.33)	6.60 (1.43)	5.53 (2.16)	5.38 (2.65)
Time (s)	134.45 (79.56)	145.67 (34.95)	159.63 (101.37)	172.30 (90.56)
Recognition test				
Time	4654.68 (1,930.37)	4026.33 (1,664.71)	4262.40 (1,769.96)	4003.41 (1,701.13)
Accuracy	82.94 (14.85)	80.80 (16.48)	82.17 (19.11)	79.55 (15.09)
Mental effort (1–9)				
Example study	5.77 (1.28)	6.07 (1.62)	5.83 (1.46)	6.14 (1.68)
Assembly test	5.03 (1.33)	6.20 (1.52)	5.63 (1.75)	6.00 (1.85)
Recognition test	6.06 (1.03)	5.67 (1.37)	6.37 (1.45)	5.72 (1.58)

instructional videos from the third-person perspective led to more errors related to the placement of components on the circuit board, but not more errors related to the specific components used to reassemble the circuit.

How do the treatments affect performance on the recognition test? A factorial ANOVA was conducted, with perspective (first-person or third-person), imitation (imitate or no-imitate), and circuit order (low-high complexity or high-low complexity) serving as between-subjects factors, and recognition test accuracy and response time serving as dependent measures. The analysis indicated no main effects of perspective on recognition test accuracy, $F(1, 97) = 2.25$, $p = .137$, or average response time, $F(1, 97) < 1$, $p = .771$. Further, there were no significant main effects of imitating on recognition test accuracy, $F(1, 97) < 1$, $p = .545$, or response time (although marginal), $F(1, 97) = 3.27$, $p = .074$. Finally, none of the other main effects or interactions among the factors were significant. Possibly the recognition test was not sensitive to the treatments in this study.

How do the treatments affect cognitive load? A mixed factorial ANOVA was conducted, with perspective (first-person or third-person), imitation (imitate or no-imitate), and circuit order (low-high complexity or high-low complexity) serving as between-subjects factors, circuit complexity (low or high) serving as a within-subjects factor, and self-reported cognitive load ratings as the dependent measures.

The analysis indicated no significant main effects of perspective on self-reported cognitive load throughout the experiment: after watching the low complexity video, $F(1, 97) < 1$, $p = .543$, assembling the low complexity circuit, $F(1, 97) = 1.96$, $p = .165$, watching the high complexity video, $F(1, 97) = 2.72$, $p = .102$, assembling the high complexity circuit, $F(1, 97) = 2.37$, $p = .127$, or completing the recognition test $F(1, 97) = 2.47$, $p = .119$. However, there were significant main effects of imitating, such that students who imitated along with the video model reported less cognitive load while watching the low complexity video, $F(1, 97) = 35.46$, $p < .001$, assembling the low complexity circuit, $F(1, 97) = 17.01$, $p < .001$, and watching the high complexity video, $F(1, 97) = 16.68$, $p < .001$. This difference did not reach statistical significance for assembling the high complexity circuit, $F(1, 97) = 3.29$, $p = .073$, and for completing the recognition test, $F(1, 97) = 2.84$, $p = .093$. There were no other significant main effects or interactions involving self-reported cognitive load. Overall, imitating along with the video model appears to reduce cognitive load while watching the video as well as while assembling the circuit.

Summary

Data from Experiment 1 provide initial evidence that students learn an assembly task better when instruction is presented from a first-person perspective rather than a third-person perspective. As expected, this effect was strongest for the high-complexity task, suggesting that the increased cognitive demands of the task make it more difficult for learners to overcome the detrimental effects of viewing the to-be-learned actions from the third-person perspective. Somewhat surprisingly, imitating the model's actions during learning did not appear to alleviate the influence of perspective on test performance. Experiment 2 aimed to determine whether it is possible to replicate the perspective findings and investigated whether explaining during test performance would moderate the detrimental effects of the third-person perspective video examples on test performance.

Experiment 2

The purpose of Experiment 2 was to replicate and extend the findings from Experiment 1 in another lab. First, we attempted to replicate the perspective effect using the high-complexity task from Experiment 1. Second, we tested whether a different type of learning strategy—informing participants that they would have to generate a verbal explanation during the test—might help compensate for viewing the instructional video from the third-person perspective. Since the verbal instructions provided by the model are spoken from the first-person perspective, we reasoned that informing learners that they would have to explain during reassembly might focus their attention on the model's explanation. Subsequently providing this explanation themselves might help them mentally transform actions observed from the third-person perspective into their own perspective during the test. Thus, Experiment 2 served to further test the generalizability and robustness of the perspective effect across learning contexts.

Method

Participants and design. The participants were 121 students, recruited from the subject pool of the behavioral lab of a Dutch university. One participant was excluded from the sample for failing to comply with the instructions during the experiment, leaving 120 participants. They were informed prior to signing up that the experiment would be conducted in English.[2] Participants gave informed consent during the process of signing up for the study via one of the online recruitment portals and participated either to fulfill a course requirement (psychology students, $n = 93$, 77.5%) or for a monetary reward of 5 Euro (approximately 5.43 USD at the time of writing). The mean age of participants was 21.97 years ($SD = 3.03$), and there were 68 women and 52 men. Participants were randomly assigned to one of four conditions, based on two between-subjects factors—perspective of the instructional videos (first-person or third-person), and whether or not students explained how to build the circuit to a fictitious other student during the building test (explaining or no-explaining). There were 30 students in the first-person/explaining group, 31 in the first-person/no-explaining group, 29 in the third-person/explaining group, and 30 in the third-person/no-explaining group. The groups did not significantly differ in terms of average age, proportion of men and women, or self-reported experience with circuits.

Materials. The paper-based demographics questionnaire and subjective mental effort scale were identical to those used in Experiment 1. The computer-based materials were identical to the no-imitation condition materials used in Experiment 1, with the exception that only the example video and test tasks for the high-complexity circuit were used in Experiment 2.

Apparatus. The apparatus consisted of two Hewlett-Packard computers with 22-in. screens, two Sennheiser PX30 headphones, and two web cameras.

[2] Note that this university has an international orientation; in most study programs the majority of the course literature is in English and lectures and work groups are also frequently in English.

Procedure. The participants were randomly assigned to conditions, and were tested in sessions of approximately 30 min., with a maximum of two participants per session. Participants were seated in a cubicle, which was equipped with a PC monitor on which the stimuli were presented and a webcam that was used to record their performance and explanation. First, participants provided informed consent and completed the demographics questionnaire. Participants in the no-explaining condition were then instructed that they would be watching a video example on how to build an electric circuit and that they would be asked to build it themselves afterward, whereas participants in the explain condition were instructed that they would be watching a video example on how to build an electric circuit and that they would be asked to demonstrate and explain to another student how to build it afterward. Then participants watched the instructional video demonstrating how to build the (high-complexity) circuit, from either the first-person or third-person perspective depending on assigned condition (and they all watched the video without imitating the video model). After watching the instructional video, participants rated how much effort they invested in studying it and then were asked to assemble the circuit on their own using the electric circuit kit (no-explaining condition) or to demonstrate and explain to another student how to build the circuit using the electric circuit kit (explaining condition). After attempting to assemble the circuit, participants rated how much effort they invested in this task. Finally, participants completed the recognition test, and rated how much effort they invested in this test.

Results and Discussion

Do students learn better from videos recorded from a first-person perspective than from a third-person perspective? Means and standard deviations for assembly test accuracy and assembly test time are shown in Table 2. All data were analyzed with 2 × 2 ANOVAs with perspective (first-person or third-person) and explanation (yes or no) as between-subjects factors, unless indicated otherwise.

In line with our main hypothesis and replicating the findings from Experiment 1, there was a significant main effect of perspective on assembly accuracy, $F(1, 116) = 10.64, p < .001, \eta_p^2 = .08$, with participants who had observed examples from the first-person perspective ($M = 6.62$, $SD = 1.37$) outperforming participants who had observed examples from the third-person perspective ($M = 5.46$, $SD = 2.39$). However, in contrast to our hypothesis and the findings from Experiment 1, the difference between the solution time of the first-person perspective group ($M = 139.97$, $SD = 61.54$) and the third-person perspective group ($M = 165.97$, $SD = 95.51$) did not reach statistical significance, $F(1, 116) = 3.12, p = .080, \eta_p^2 = .03$, in Experiment 2. Overall, there is partial support for a replication of the perspective effect found in Experiment 1.

Does the perspective effect depend on whether students explained what they were doing on the assembly test? Explanation instructions did not improve accuracy on the assembly test, $F(1, 116) < 1, p = .781$, and did not compensate for third-person perspective effects, as there was no significant interaction between perspective and explanation, $F(1, 116) < 1, p = .879$. Similarly, for assembly time, there was no main effect of explanation, $F(1, 116) < 1, p = .417$, nor an interaction effect between explanation and perspective, $F(1, 116) < 1, p = .961$. Overall, it appears that explaining during test performance was not a boundary condition (or moderator) for the perspective effect, and was not an effective technique for improving performance.

Does perspective affect the type of errors students make on assembly tasks? As in Experiment 1, we analyzed whether there were differences between the first-person and third-person perspective conditions in the number of students who made perspective errors, location errors, and component errors during the assembly test, using chi-square tests. In the third-person perspective condition, more students (34 out of 59, or 57.6%) made a perspective error than in the first-person condition (0 out of 61, or 0%), $\chi^2(1) = 49.05, p < .001$ (2-sided). However, there were no differences in the number of students who made location errors (in contrast to Experiment 1), $\chi^2(1) < 1, p = .594$ (first-person: 13/61 or 21.3%; third-person: 15/59, or 25.4%), or component errors (in line with Experiment 1), $\chi^2(1) < 1, p = .946$ (first-person: 41/61, or 67.2%; third-person: 40/59, or 67.8%).

How do the treatments affect recognition test performance? As in Experiment 1, there were no main effects of perspective on recognition test accuracy, $F(1, 116) < 1, p = .738$, or average response time on the correct trials, $F(1, 116) < 1, p = .552$. Further, there was no significant main effect of explaining on recognition test accuracy, $F(1, 116) < 1, p = .431$, or response time on the correct trials, $F(1, 116) = 1.88, p = .173$, and no significant interaction on recognition test accuracy, $F(1, 116) < 1, p = .937$, or response time on the correct trials, $F(1, 116) < 1, p = .569$.

How do the treatments affect cognitive load? Analysis of the self-reported mental effort invested in example study showed no main effect of perspective, $F(1, 116) < 1, p = .814$, or explaining, $F(1, 116) = 1.16, p = .283$, nor an interaction effect, $F(1, 116) < 1, p = .983$. The analysis of mental effort invested in the assembly test, did show a main effect of explaining, $F(1, 116) = 6.72, p = .011, \eta_p^2 = .05$, indicating—as one would expect—that participants who explained during the assembly test ($M = 6.10$, $SD = 1.68$) reported higher effort than participants who did not explain ($M = 5.33$, $SD = 1.57$). There was no main effect of perspective on effort invested in the assembly test, $F(1, 116) < 1, p = .499$, nor an interaction between perspective and explaining, $F(1, 116) = 1.83, p = .179$.

On the recognition test, explaining seemed to have a significant effect in the opposite direction: participants who had explained on the assembly test, reported lower effort investment on the recognition test ($M = 5.69$, $SD = 1.47$) than participants who had not explained ($M = 6.21$, $SD = 1.25$), $F(1, 116) = 4.33, p = .040, \eta_p^2 = .04$; however, since the test of the overall model was not significant, this effect should be interpreted with caution. There was no main effect of perspective on effort invested in the recognition test, $F(1, 116) < 1, p = .473$, nor an interaction between perspective and explaining, $F(1, 116) < 1, p = .625$. Overall, the recognition test may not be a sensitive measure.

General Discussion

Empirical Contributions

Across two experiments conducted in different labs, students learned better from instructional videos recorded from a first-

person perspective than a third-person perspective as indicated by better accuracy (significant with a medium effect size in Experiment 1, $\eta_p^2 = .06$, and Experiment 2, $\eta_p^2 = .08$) and faster solution time (significant in Experiment 1, $\eta^2 = .06$, but not significant in Experiment 2, $\eta_p^2 = .02$) on an assembly test. An important boundary condition identified in Experiment 1 is that the perspective effect was strong for high-complexity assembly tasks but not for low-complexity assembly tasks. The effect of perspective on complex tasks was found whether or not students imitated the instructor during learning (in Experiment 1) and whether or not students engaged in explaining during the assembly test (in Experiment 2). Overall, this study extends basic empirical research on the facilitative role of a first-person perspective viewpoint to the design of video modeling examples of an assembly task.

Theoretical Contributions

The present study was based on predictions from embodied theories of cognition, which posit that human thought and action is deeply grounded in one's personal sensory-motor experiences of the physical world (Barsalou, 2008; Wilson, 2002). Accordingly, the first-person perspective is assumed to be critical in shaping one's internal representations of observed spatial relations and actions. We predicted that observing to-be-performed actions from the first-person perspective would facilitate the construction of a more accurate mental representation of those actions, and result in better subsequent performance, compared to observing to-be-performed actions from the third-person perspective. Findings from both experiments provided support for this prediction, demonstrating that students were generally more accurate, faster, and made fewer errors on an assembly task after viewing instructional videos presented in the first-person perspective.

Furthermore, Experiment 1 supported our second prediction that the perspective effect would be strongest for the high-complexity task compared to the low-complexity task. Observing to-be-performed actions from the third-person perspective presumably requires students to generate additional visual-spatial transformations to convert the information into their own perspective. Tasks relatively low in complexity do not excessively tax learners' limited working memory resources, leaving ample resources for making such mental transformations. However, this is different for tasks relatively high in complexity, which require students to represent a greater number of interacting elements in working memory. On such complex tasks, having to make these transformations may overload students' processing capacity and result in impaired learning.

The present study did not support predictions that imitation during example study (Experiment 1) and explaining during the building test (Experiment 2) would compensate for the negative effect of the third-person perspective on test performance. Experiment 1 indicated that imitation led to faster assembly time and reduced subjective reports of invested mental effort, but did not influence assembly accuracy. This suggests that imitation may have led to a practice effect that increased efficiency during subsequent assembly, although it did not improve test performance. This lack of effect on test performance is interesting in light of other studies that have shown that—at least in relatively short learning phases—example study followed by practice problem solving is not more effective than example study only (Leahy, Hanham, & Sweller, 2015; van Gog & Kester, 2012; van Gog et al., 2015), even when the practice opportunity is additional (i.e., not replacing an example study opportunity; Baars, van Gog, De Bruin, & Paas, 2014). This seems to underline the notion that imitation is not strictly necessary for observational learning to occur (Bandura, 1986), but note that imitation may become important in longer training sessions, to refine and automate performance (for which imitation is effective, as shown here by the reduced time on task and effort investment). More importantly in light of our present study, the experience of imitating a video model did not appear to help students overcome the detrimental effects on test performance of observing the video from the third-person perspective compared to the first-person perspective.

Similarly, in Experiment 2, knowing that one had to explain during the assembly test and actually giving an explanation to a (fictitious, nonpresent) other student, was not enough to boost performance generally (i.e., in both perspective conditions), or to counteract the cognitive demands of observing to-be-performed actions from the third-person perspective. Previous research has shown—in line with findings on self-explaining (Wylie & Chi, 2014) and peer tutoring (Roscoe & Chi, 2007)—that explaining to fictitious, nonpresent others on video camera improves learning, as evidenced by later test performance (Fiorella & Mayer, 2013, 2014; Hoogerheide, Loyens, & van Gog, 2014; Hoogerheide, Deijkers, Loyens, Heijltjes, & van Gog, 2016). In the present study, however, we investigated effects of explaining during the test itself. It is possible that the beneficial effects on knowledge restructuring that are often found to result from explaining, only manifest themselves at a later point in time.

Practical Contributions

Although the use of video modeling examples within formal and informal settings is growing rapidly, there is a paucity of rigorous empirical research to inform educators and instructional designs on how to design video lessons effectively. The present study provides preliminary evidence for a design principle of instructional videos that can be called the *perspective principle*: people learn better when instructional videos are recorded from a first-person perspective rather than a third-person perspective. This principle appears to apply most strongly for videos depicting complex tasks. Whether it applies only to learning from modeling examples on manual assembly tasks, such as assembling the components of an electric circuit, or also to other types of modeling examples or video instructions, is a question for future research to address. Our study also shows that the perspective effect is not remedied by engaging in generative learning strategies such as imitation and explaining. Although we cannot rule out the possibility that other strategies might be more successful, this strongly suggests that it is better to prevent using third-person perspective videos and create first-person videos whenever possible.[6]

The practical relevance of our findings is strengthened by the fact that we replicated the perspective effect in two different labs (in different nations with different subject populations), indicating that the perspective effect is robust. This cross-national collaboration reflects the idea that replication is a crucial aspect of educational research (Makel & Plucker, 2014; Shavelson & Towne, 2002). Overall, the findings suggest that video modeling

examples can be enhanced by presenting to-be-performed actions from the learner's own perspective.

Limitations and Future Directions

This study involved a short video on a single topic presented in a lab environment with learning assessed on an immediate test. Further work is needed to determine whether the perspective effect can be found with videos on other topics, in authentic educational settings, and on delayed tests. The perspective effect may be applicable across a wide range of domains, such as other assembly tasks, and more broadly, other types of complex motor tasks, such as when learning movements in sports, dance, or in playing a musical instrument. In academic learning, the perspective effect may extend to teaching topics in STEM domains, such as in the use of physical and virtual models to teach complex spatial relations of molecules in chemistry, or in the use of concrete manipulatives teach abstract math concepts. It may also apply to viewing other types of instructor movements, such as by viewing gestures or drawings from the first-person perspective, or to other lesson formats, such as a series of static images of an instructor manipulating objects. It might also be interesting to investigate whether the perspective effect would be stronger when viewing dynamic videos than when observing static images, as the videos might result in stronger motor-neuron activation.

The present study generally did not find effects of perspective on the recognition test or on subjective mental effort, and no complexity effect on effort in Experiment 1. The recognition test may not have been sensitive enough for detecting an influence of viewing materials from different perspectives, given that recognizing completed circuits is somewhat distinct from building a circuit on one's own, and performance was quite high in all conditions. One possibility is that the information from the instructional videos is represented in memory as actions, and therefore the recognition test did not capture learners' action-based representations.

The subjective mental effort ratings capture the overall cognitive load experienced by a learner while viewing an example or completing a task (Paas, Tuovinen, Tabbers, & Van Gerven, 2003). Given that we see all kinds of tasks being performed from a third-person perspective on a daily basis, it is not that surprising that learners do not experience higher cognitive load when studying third-person examples. On the building test, however, one might expect that having to translate from the observed third-person perspective to a first-person perspective would impose higher cognitive load and require more effort. One drawback of the fact that overall load is measured, however, is that we do not know from which cognitive processes it originates; participants in two different conditions can experience the same amount of cognitive load, while the processes from which it originates can be beneficial for learning or performance in one condition (as evidenced by higher test performance) and detrimental for learning in the other condition (as evidenced by lower test performance). We can conclude though, that the first-person perspective was more efficient, given that better test performance is reached with similar levels of effort invested in example study and test performance (van Gog & Paas, 2008). With regard to task complexity in Experiment 1, the effort ratings were somewhat higher for the more complex task, but not significantly higher. Possibly, this is due to the fact that both tasks required eight steps to be memorized. In future research, continuous and more objective measures of cognitive load, such as dual-task measures (Brünken, Plass, & Leutner, 2003) or physiological measures such as EEG (Antonenko, Paas, Grabner, & van Gog, 2010) or eye tracking (van Gog & Jarodzka, 2013) may help attain more insight into cognitive processing demands associated with viewing video examples from first- or third-person perspective.

Research should also continue to explore potential boundary conditions and moderating factors of the perspective effect. For example, individual differences such as prior knowledge, spatial ability, and working memory capacity may moderate the benefits of viewing instructional videos from the first-person perspective. That is, the perspective effect may be strongest for learners with low prior knowledge, low spatial ability, or low working memory capacity, because they do not have sufficient cognitive capacity to mentally represent and convert actions from the third-person perspective to their own perspective. The current study included a measure of spatial perspective taking; however, it was not predictive of performance on the assembly task, and so could not be explored as a potential moderator. Testing the perspective effect with more content-rich materials would also allow researchers to better explore the role of learners' prior knowledge in learning from first- and third-person perspective. In short, research is needed to clarify the generalizability of the perspective effect within different educational contexts.

Conclusion

Overall, this study contributes toward a theoretical understanding of how students learn from instructional videos and a practical understanding of how to help students learn from instructional videos. Presenting instructional videos from the learner's perspective (as opposed to from a third-person perspective) appears to better support the construction of appropriate visuospatial representations during learning, thereby resulting in better subsequent task performance. As the use of video in education continues to accelerate, this basic empirical finding offers important implications for instructional design, potentially applicable across a wide range of learning environments.

References

Antonenko, P., Paas, F., Grabner, R., & van Gog, T. (2010). Using electroencephalography (EEG) to measure cognitive load. *Educational Psychology Review, 22,* 425–438. http://dx.doi.org/10.1007/s10648-010-9130-y

Arguel, A., & Jamet, E. (2009). Using video and static pictures to improve learning of procedural contents. *Computers in Human Behavior, 25,* 354–359. http://dx.doi.org/10.1016/j.chb.2008.12.014

Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction, 16,* 389–400. http://dx.doi.org/10.1016/j.learninstruc.2006.09.001

Ayres, P., Marcus, N., Chan, C., & Qian, N. (2009). Learning hand manipulative tasks: When instructional animations are superior to equivalent static representations. *Computers in Human Behavior, 25,* 348–353. http://dx.doi.org/10.1016/j.chb.2008.12.013

Baars, M., van Gog, T., De Bruin, A., & Paas, F. (2014). Effects of problem solving after worked example study on primary school children's monitoring accuracy. *Applied Cognitive Psychology, 28,* 382–391. http://dx.doi.org/10.1002/acp.3008

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review, 84,* 191–215.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59,* 617–645. http://dx.doi.org/10.1146/annurev.psych.59.103006.093639

Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38,* 53–61. http://dx.doi.org/10.1207/S15326985EP3801_7

Castro-Alonso, J. C., Ayres, P., & Paas, F. (2015). Animations showing Lego manipulative tasks: Three potential moderators of effectiveness. *Computers & Education, 85,* 1–13. http://dx.doi.org/10.1016/j.compedu.2014.12.022

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4–58. http://dx.doi.org/10.1177/1529100612453266

Fiorella, L., & Mayer, R. E. (2013). The relative benefits of learning by teaching and teaching expectancy. *Contemporary Educational Psychology, 38,* 281–288. http://dx.doi.org/10.1016/j.cedpsych.2013.06.001

Fiorella, L., & Mayer, R. E. (2014). Role of expectations and explanations in learning by teaching. *Contemporary Educational Psychology, 39,* 75–85. http://dx.doi.org/10.1016/j.cedpsych.2014.01.001

Fiorella, L., & Mayer, R. E. (2015a). Eight ways to promote generative learning. *Educational Psychology Review.* Advance online publication. http://dx.doi.org/10.1007/s10648-015-9348-9

Fiorella, L., & Mayer, R. E. (2015b). *Learning as a generative activity.* New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781107707085

Fiorella, L., & Mayer, R. E. (2016). Effects of observing the instructor draw diagrams on learning from multimedia messages. *Journal of Educational Psychology, 108,* 528–546. http://dx.doi.org/10.1037/edu0000065

Hegarty, M., & Waller, D. (2004). A dissociation between mental rotation and perspective-taking spatial abilities. *Intelligence, 32,* 175–191. http://dx.doi.org/10.1016/j.intell.2003.12.001

Hoogerheide, V., Deijkers, L., Loyens, S. M. M., Heijltjes, A., & van Gog, T. (2016). Gaining from explaining: Learning improves from explaining to fictitious others on video, not from writing to them. *Contemporary Educational Psychology, 44/45,* 95–106. http://dx.doi.org/10.1016/j.cedpsych.2016.02.005

Hoogerheide, V., Loyens, S. M. M., & van Gog, T. (2014). Effects of creating video-based modeling examples on learning and transfer. *Learning and Instruction, 33,* 108–119. http://dx.doi.org/10.1016/j.learninstruc.2014.04.005

Hoogerheide, V., Loyens, S. M. M., & van Gog, T. (2016). Learning from video modeling examples: Does gender matter? *Instructional Science, 44,* 69–86. http://dx.doi.org/10.1007/s11251-015-9360-y

Hoogerheide, V., van Wermeskerken, M., Loyens, S. M. M., & van Gog, T. (2016). Learning from video modeling examples: Content kept equal, adults are more effective models than peers. *Learning and Instruction, 44,* 22–30. http://dx.doi.org/10.1016/j.learninstruc.2016.02.004

Jackson, P. L., Meltzoff, A. N., & Decety, J. (2006). Neural circuits involved in imitation and perspective-taking. *NeuroImage, 31,* 429–439. http://dx.doi.org/10.1016/j.neuroimage.2005.11.026

Kelly, R. L., & Wheaton, L. A. (2013). Differential mechanisms of action understanding in left and right handed subjects: The role of perspective and handedness. *Frontiers in Psychology, 4,* 957. http://dx.doi.org/10.3389/fpsyg.2013.00957

Kessler, K., & Thomson, L. A. (2010). The embodied nature of spatial perspective taking: Embodied transformation versus sensorimotor interference. *Cognition, 114,* 72–88. http://dx.doi.org/10.1016/j.cognition.2009.08.015

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic action from pragmatic action. *Cognitive Science, 18,* 513–549. http://dx.doi.org/10.1207/s15516709cog1804_1

Kizilcec, R. F., Bailenson, J. N., & Gomez, C. J. (2015). The instructor's face in video instruction: Evidence from two large-scale field studies. *Journal of Educational Psychology, 107,* 724–739. http://dx.doi.org/10.1037/edu0000013

Kockler, H., Scheef, L., Tepest, R., David, N., Bewernick, B. H., Newen, A., . . . Vogeley, K. (2010). Visuospatial perspective taking in a dynamic environment: Perceiving moving objects from a first-person-perspective induces a disposition to act. *Consciousness and Cognition, 19,* 690–701. http://dx.doi.org/10.1016/j.concog.2010.03.003

Kozhevnikov, M., Motes, M. A., Rasch, B., & Blajenkova, O. (2006). Perspective-taking vs. mental rotation transformations and how they predict spatial navigation performance. *Applied Cognitive Psychology, 20,* 397–417. http://dx.doi.org/10.1002/acp.1192

Leahy, W., Hanham, J., & Sweller, J. (2015). High element interactivity information during problem solving may lead to failure to obtain the testing effect. *Educational Psychology Review, 27,* 291–304. http://dx.doi.org/10.1007/s10648-015-9296-4

Lindgren, R. (2012). Generating a learning stance through perspective-taking in a virtual environment. *Computers in Human Behavior, 28,* 1130–1139. http://dx.doi.org/10.1016/j.chb.2012.01.021

Lorey, B., Bischoff, M., Pilgramm, S., Stark, R., Munzert, J., & Zentgraf, K. (2009). The embodied nature of motor imagery: The influence of posture and perspective. *Experimental Brain Research, 194,* 233–243. http://dx.doi.org/10.1007/s00221-008-1693-1

Maeda, F., Kleiner-Fisman, G., & Pascual-Leone, A. (2002). Motor facilitation while observing hand actions: Specificity of the effect and role of observer's orientation. *Journal of Neurophysiology, 87,* 1329–1335.

Makel, M. C., & Plucker, J. A. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43,* 304–316. http://dx.doi.org/10.3102/0013189X14545513

Marcus, N., Cleary, B., Wong, A., & Ayres, P. (2013). Should hand actions be observed when learning hand motor skills from instructional animations? *Computers in Human Behavior, 29,* 2172–2178. http://dx.doi.org/10.1016/j.chb.2013.04.035

Marley, S. C., & Carbonneau, K. J. (2015). How psychological research with instructional manipulatives can inform classroom learning. *Scholarship of Teaching and Learning in Psychology, 1,* 412–424. http://dx.doi.org/10.1037/stl0000047

Ouwehand, K., van Gog, T., & Paas, F. (2015). Designing effective video-based modeling examples using gaze and gesture cues. *Journal of Educational Technology & Society, 18,* 78–88.

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive load approach. *Journal of Educational Psychology, 84,* 429–434. http://dx.doi.org/10.1037/0022-0663.84.4.429

Paas, F., & Sweller, J. (2012). An evolutionary upgrade of cognitive load theory: Using the human motor system and collaboration to support the learning of complex cognitive tasks. *Educational Psychology Review, 24,* 27–45. http://dx.doi.org/10.1007/s10648-011-9179-2

Paas, F., Tuovinen, J., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38,* 63–71. http://dx.doi.org/10.1207/S15326985EP3801_8

Paas, F., & Van Merrienboer, J. J. G. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6,* 351–371. http://dx.doi.org/10.1007/BF02213420

Richardson, A. E., Montello, D. R., & Hegarty, M. (1999). Spatial knowledge acquisition from maps and from navigation in real and virtual environments. *Memory & Cognition, 27,* 741–750. http://dx.doi.org/10.3758/BF03211566

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27,* 169–192. http://dx.doi.org/10.1146/annurev.neuro.27.070203.144230

Roscoe, R. D., & Chi, M. T. H. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research, 77,* 534–574. http://dx.doi.org/10.3102/0034654307309920

Schunk, D. H., Hanson, A. R., & Cox, P. D. (1987). Peer-model attributes and children's achievement behaviors. *Journal of Educational Psychology, 79,* 54–61. http://dx.doi.org/10.1037/0022-0663.79.1.54

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific research in education.* Washington, DC: National Academy Press.

Surtees, A. D., & Apperly, I. A. (2012). Egocentrism and automatic perspective taking in children and adults. *Child Development, 83,* 452–460.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory.* New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4419-8126-4

Sweller, J., & Sweller, S. (2006). Natural information processing systems. *Evolutionary Psychology, 4,* 434–458. http://dx.doi.org/10.1177/147470490600400135

van Gog, T., & Jarodzka, H. (2013). Eye tracking as a tool to study and enhance cognitive and metacognitive processes in computer-based learning environments. In R. Azevedo & V. Aleven (Eds.), *International handbook of metacognition and learning technologies* (pp. 143–156). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4419-5546-3_10

van Gog, T., & Kester, L. (2012). A test of the testing effect: Acquiring problem-solving skills from worked examples. *Cognitive Science, 36,* 1532–1541. http://dx.doi.org/10.1111/cogs.12002

van Gog, T., Kester, L., Dirkx, K., Hoogerheide, V., Boerboom, J., & Verkoeijen, P. P. J. L. (2015). Testing after worked example study does not enhance delayed problem-solving performance compared to restudy. *Educational Psychology Review, 27,* 265–289. http://dx.doi.org/10.1007/s10648-015-9297-3

van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist, 43,* 16–26. http://dx.doi.org/10.1080/00461520701756248

van Gog, T., Paas, F., Marcus, N., Ayres, P., & Sweller, J. (2009). The mirror neuron system and observational learning: Implications for the effectiveness of dynamic visualizations. *Educational Psychology Review, 21,* 21–30. http://dx.doi.org/10.1007/s10648-008-9094-3

van Gog, T., & Rummel, N. (2010). Example-based learning: Integrating cognitive and social-cognitive research perspectives. *Educational Psychology Review, 22,* 155–174. http://dx.doi.org/10.1007/s10648-010-9134-7

van Gog, T., Verveer, I., & Verveer, L. (2014). Learning from video modeling examples: Effects of seeing the human model's face. *Computers & Education, 72,* 323–327. http://dx.doi.org/10.1016/j.compedu.2013.12.004

Vogeley, K., & Fink, G. R. (2003). Neural correlates of the first-person-perspective. *Trends in Cognitive Sciences, 7,* 38–42. http://dx.doi.org/10.1016/S1364-6613(02)00003-7

Vogt, S., Taylor, P., & Hopkins, B. (2003). Visuomotor priming by pictures of hand postures: Perspective matters. *Neuropsychologia, 41,* 941–951. http://dx.doi.org/10.1016/S0028-3932(02)00319-6

Watanabe, R., Higuchi, T., & Kikuchi, Y. (2013). Imitation behavior is sensitive to visual perspective of the model: An fMRI study. *Experimental Brain Research, 228,* 161–171. http://dx.doi.org/10.1007/s00221-013-3548-7

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review, 9,* 625–636. http://dx.doi.org/10.3758/BF03196322

Wylie, R., & Chi, M. T. H. (2014). The self-explanation principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 413–432). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139547369.021

Received May 17, 2016
Revision received September 14, 2016
Accepted September 15, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 5, 666–679

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000167

Can Collaborative Learning Improve the Effectiveness of Worked Examples in Learning Mathematics?

Endah Retnowati, Paul Ayres, and John Sweller
University of New South Wales

Worked examples and collaborative learning have both been shown to facilitate learning. However, the testing of both strategies almost exclusively has been conducted independently of each other. The main aim of the current study was to examine interactions between these 2 strategies. Two experiments (N = 182 and N = 122) were conducted with Grade-7 Indonesian students, comparing learning to solve algebra problems, with higher and lower levels of complexity, collaboratively or individually. Results from both experiments indicated that individual learning was superior to collaborative learning when using worked examples. In contrast, in Experiment 2, when learning from problem solving using problem-solving search, collaboration was more effective than individual learning. However, again in Experiment 2, studying worked examples was overall superior to learning from solving problems, particularly for more complex problems. It can be concluded that while collaboration could be beneficial when learning under problem solving conditions, it may be counterproductive when studying worked examples.

Keywords: worked examples, cognitive load theory, collaboration, problem complexity

Across multiple domains ranging from mathematics to visual arts, researchers have demonstrated that when learning novel material, guided instruction through worked examples is more effective for novice learners than conventional problem solving strategies (see Atkinson, Derry, Renkl, & Wortham, 2000; P. A. Kirschner, Sweller, & Clark, 2006; Renkl, 2014a, 2014b; Sweller, Ayres, & Kalyuga, 2011). However, most of the research into worked examples has focused exclusively on individual learning settings. Few attempts have investigated worked examples in collaborative settings (e.g., F. Kirschner, Paas, Kirschner, & Janssen, 2011; Retnowati, Ayres, & Sweller, 2010).

A main aim of the current study (Experiment 1) was to investigate the effectiveness of collaborative learning compared with individual learning within a worked examples environment. Another aim (Experiment 2) was to compare possible interactions between studying individually or collaboratively on the one hand and studying worked examples or solving problems on the other hand. We will begin by outlining the worked example effect.

The Worked Example Effect

A worked example provides a step-by-step solution to a problem or task and is a form of explicit instruction (see P. A. Kirschner et al., 2006). Rather than trying to acquire new information through problem-solving search or other types of discovery methods, learners are shown worked examples to study. Worked examples provide an expert's problem-solving model, from which students can study and learn (Atkinson et al., 2000). With worked examples, learners are able to focus on understanding a solution rather than focus on solving the problem (Renkl, 2014a). The worked example effect occurs when students who learn from studying worked examples subsequently obtain superior test scores to students who learn from solving problems. From this perspective, we refer to problem solving as solving problems with minimal teacher/instructor guidance on how to solve the problem.

Using algebra problems, Cooper and Sweller (1987) and Sweller and Cooper (1985) provided the first demonstrations of the worked example effect (Sweller et al., 2011). They found that students who were asked to study worked examples performed better on subsequent problem solving tests than students required to practice solving the equivalent problems. The effect was explained by the suggestion that worked examples reduced extraneous working memory load compared to solving the equivalent problems. A reduction in cognitive load facilitated the transfer of knowledge to long-term memory. These findings led to further research in mathematics and scientific domains. For example, the worked example effect was replicated in algebra (Carroll, 1994), geometry (Paas & van Merriënboer, 1994; Tarmizi & Sweller, 1988), statistics (Paas, 1992; Quilici & Mayer, 1996), and physics (Ward & Sweller, 1990) using a range of age groups and subject areas with the advantage appearing for both similar and transfer problems. Building on these initial findings, more contemporary research (for summaries, see Ayres & Sweller, 2013; Renkl, 2014a) has found the effect in nonscience domains such as visual arts (Rourke & Sweller, 2009) and English literature (Kyun, Kalyuga, & Sweller, 2013; Oksa, Kalyuga, & Chandler, 2010), as well as ongoing investigations in the science domain, such as problem solving in electrical circuits (van Gog & Kester, 2012), and geometry (Chen, Kalyuga, & Sweller, 2015, 2016a).

This article was published Online First December 19, 2016.

Endah Retnowati, Paul Ayres, and John Sweller, School of Education, University of New South Wales.

Endah Retnowati is now at the Department of Mathematics, Education Faculty of Mathematics and Sciences, Yogyakarta State University, Indonesia.

Correspondence concerning this article should be addressed to Paul Ayres, School of Education, University of New South Wales, Sydney, NSW 2052, Australia. E-mail: p.ayres@unsw.edu.au

Worked examples illustrate one of the main principles of cognitive load theory, the borrowing and reorganizing principle (see Sweller & Sweller, 2006). This principle suggests that the most effective way to obtain new information is by directly receiving it from another person who already has this information. The major mechanisms are listening to other people, reading what they write, and imitating what they do. In that sense, information is borrowed from another person's long-term memory. In the case of worked examples, information is borrowed from the long-term memory of the constructor of the worked examples. However, this information is reorganized by the learner by integrating the new information with old information stored in the learner's long-term memory (see Sweller et al., 2011). Integrating new information with old information that is already understood may assist in making sense of the new information. This conceptualization is consistent with other theories of learning that emphasize reorganization and elaboration (see Mayer, 2014). Because worked examples provide a low cognitive load environment compared with problem solving search, learning is enhanced through the construction of new schematic knowledge.

Virtually all published research into worked examples has been conducted using individual rather than collaborative learning. A notable exception was a study by Retnowati et al. (2010), who found in a single experiment using Grade 7 Indonesian students that worked examples in geometry were superior to problem solving for both individual and collaborative learners on both retention and transfer tasks that required calculations as well as providing related explanations. Qualitative data also revealed that participants believed that they understood the material more easily when using worked examples rather than problem solving. A goal of the present study was to investigate whether worked examples could be enhanced by using collaborative settings.

Collaborative Learning

Collaborative learning occurs when students learn by collaborating rather than by studying individually. It is widely used (Gillies, 2003) and considered highly desirable in the community and workplace (Barron, 2000). Considerable evidence suggests that collaborative learning has significant academic, social, and psychological benefits (Johnson, Johnson, & Smith, 1998). Multiple studies and meta-analyses have found that the various forms of collaborative or cooperative learning strategies where students work together have significant benefits over students who work individually (see Johnson, Maruyama, Johnson, Nelson, & Skon, 1981). Many of these studies have focused on learning mathematics, showing that small-group learning has led to greater mathematical outcomes than traditional methods of teaching individuals (see Davidson & Kroll, 1991). Explanations for this advantage are usually grounded in social constructivist theory or social independence theory, which emphasize that learning should be facilitated through social and collaborative activities where students construct knowledge by interactions with others and through collective goals (Johnson & Johnson, 1994; Schreiber & Valle, 2013).

Studies have been conducted to identify the factors that improve collaborative learning (for reviews, see Cohen, 1994; Kreijns, Kirschner, & Jochems, 2003; Schreiber & Valle, 2013; Van den Bossche, Gijselaers, Segers, & Kirschner, 2006; Webb, 2009; Weinberger, Stegmann, & Fischer, 2007). It is generally agreed that collaborative learning requires active social interactions, group goals, and individual accountability (see Slavin, 1995).

The use of problem solving activities within collaborative learning classrooms has been strongly advocated, especially by mathematics educators (see, e.g., the National Council of Teachers of Mathematics, 2000). According to De Corte (2004), one view of mathematics learning is that it is a social construction of knowledge through collaboration. An emphasis should be placed on problem solving, reasoning, and communication, forming communities of mathematical inquiry (Goos, 2004; Staples, 2007). Shared meanings of the main concepts emerge through the interactions associated with group problem solving (Plass et al., 2013), as well as learners constructing their own ideas and individual insights (Yackel, Cobb, & Wood, 1991).

Collaborative Learning and Evolutionary Psychology

Evolutionary psychology is used as a base for cognitive load theory, and this view of cognition can be used to provide a new perspective on some of the fundamental underpinnings of collaborative learning (Paas & Sweller, 2012; Sweller et al., 2011). A key aspect of this argument comes from the work of Geary (1995, 2008, 2012), who distinguished between two types of knowledge: biologically primary and secondary knowledge. Biologically primary knowledge is knowledge that we have evolved to acquire over many generations. It is easily and unconsciously acquired and is modular with different skills likely to have evolved during different evolutionary epochs. Examples are learning to listen, speak, recognize faces. and use general problem solving strategies. Biologically secondary knowledge is knowledge that we need to acquire for cultural reasons. We have evolved to acquire secondary knowledge as a general skill. We have not evolved to acquire particular types of secondary knowledge in the same way that we have evolved to acquire particular types of primary knowledge. Virtually every topic taught in education and training establishments provides an example of biologically secondary knowledge.

Geary argued that working in a collaborative environment may be natural and effortless, because it is a biologically primary activity that humans have evolved to engage in (Geary, 1995, 2008). However, this advantage may come at a cost (Geary, 1995, 2008), as during collaborative learning, students may tend to automatically develop their general communication and coordination skills, rather than allocating more attention to the assigned biologically secondary knowledge. While Geary (2008) acknowledges that social context and interaction with teachers and peers contribute to a student's learning, he also questions whether students can learn better in social contexts, rather than through explicit instruction.

As shown in many studies (see Johnson et al., 1998), social skills may automatically be improved through collaboration, which is consistent with Geary's argument outlined above. However, learning the content of a collaborative lesson is another matter because that content most likely requires the acquisition of biologically secondary skills (e.g., mathematics) that require conscious effort. As Geary suggested, collaboration may not necessarily produce advantages in academic outcomes if no more than an automatic improvement in collaborative skills occurs.

Collaborative Learning and Cognitive Load Theory

Similar to worked examples (as argued above), collaborative learning demonstrates another example of the borrowing and reorganizing principle (see Paas & Sweller, 2012). Knowledge can be borrowed from other members of the group, and reorganized, linking new knowledge with old knowledge stored in long-term memory. Group interactions can help individuals make sense of the information and steer the reorganization of the information accordingly (see De Corte, 2004; Plass et al., 2013). Because humans have evolved to communicate, to share, and to obtain information from each other as biologically primary skills, collaborative learning may have an advantage over individual learning in that it involves sharing information and learning from each other, as occurs in everyday life (Sweller et al., 2011).

Another advantage of collaborative learning is that it may assist in learning complex materials. Complex materials are difficult to learn because they impose a heavy working memory load (Sweller et al., 2011). However, if the learning material is shared among several group members, an individual is required to process less task-relevant information, potentially reducing working memory load (F. Kirschner, Paas, & Kirschner, 2009a). Working memory resources then can be allocated to learning about important aspects of the materials by processing relevant information communicated from other group members. Based on this view, collaboration should be effective by providing group members with information that they otherwise would need to search for themselves. This potential provision of information should reduce extraneous cognitive load. In this sense, a biologically primary activity, collaboration, may provide an advantage in acquiring biologically secondary knowledge such as mathematics. Combining the limited working memory resources of several individuals should increase the resources available to all in a manner that does not occur when students are engaged in individual learning and have to deal with all the working memory load themselves. Hence, through collaboration, individuals may be better able to learn about complex materials.

Initial experimental evidence in support of this hypothesis was found by F. Kirschner, Paas, and Kirschner (2009b) using a high-school biology topic, where an individual learning condition was compared to a collaborative learning condition consisting of three group members. During the learning phase, students were given problem-solving tasks to complete individually or collaboratively. For the collaborative learning condition, every member of a group had information about one third of the whole task only, and hence sharing was required to complete the task. In the individual condition, individual students were given the whole task to solve. Following the learning phase, all students were tested individually using retention and transfer tasks. Significant interaction effects were found. For the retention tasks individuals learned more efficiently, while for the transfer tasks collaboration led to more efficient learning. In a follow up study also with biology content, F. Kirschner et al. (2011) found that collaborative learning was more effective than individual learning on high but not low complexity tasks.

The Current Study

The evidence described so far suggests that both worked examples and collaborative learning are effective learning strategies. This study aimed to extend the research into both strategies by combining them in an authentic learning environment. More specifically, the main research question was to investigate if the effectiveness of worked examples could be improved by using collaborative learning. If, as indicated above, the borrowing and reorganizing principle suggests that most learning is based on obtaining information from others, then the use of collaboration permits learners to not only obtain information from explicit instruction via worked examples, but also obtain information from co-learners. By studying worked examples collaboratively, learners may obtain additional information from other learners that would not be available if learning individually.

Notwithstanding the possible advantage of adding information from collaborators to the information obtained from a worked example, the expertise reversal effect suggests that for given levels of expertise and complexity (Chen, Kalyuga, & Sweller, 2016b), the provision of additional information can become redundant, resulting in an increase rather than a decrease in cognitive load (Kalyuga, Ayres, Chandler, & Sweller, 2003). Evidence for such an outcome was obtained by Nihalani, Mayrath, and Robinson (2011). They found that for novices, feedback was more effective than collaboration. For more expert learners, the addition of feedback reduced learning and reduced the advantages of expertise. For these learners, feedback was redundant and redundancy has been shown repeatedly to interfere with learning due to an increased extraneous cognitive load. Hence, there may be conditions under which the combination of collaboration and worked examples may be less advantageous.

We also investigated how problem complexity has an impact on the effectiveness of collaborative learning and worked examples. Furthermore, because of the design of the experiments it was possible under the conditions to examine if the collaborative context was superior to individual learning, and whether worked examples were superior to problem solving. Throughout the study, authentic classroom environments were used rather than laboratory conditions.

Study Hypotheses

Hypothesis 1: Worked examples will be enhanced by studying collaboratively compared to studying individually. This hypothesis was based on research that argues collaborative learning is superior to individual learning (see Johnson & Johnson, 2002). In addition considerations of evolutionary psychology suggest that humans have evolved to collaborate naturally (Geary, 1995, 2008) and that collaborative environments provide an effective way to obtain new information by directly receiving it from another person who already has this information (Paas & Sweller, 2012; Sweller & S. Sweller, 2006).

Hypothesis 2. The effectiveness of collaborative learning will be increased by task complexity. This hypothesis flows from the general research into collaborative learning and problem complexity (see Cohen, 1994). Evidence for the impact of complex tasks in collaborative learning compared to individual learning has been demonstrated by F. Kirschner, Paas, and Kirschner (2011) using biology content, and by Zhang, Ayres, and Chan (2011) using web design materials. As reported, F. Kirschner et al. (2011) also showed that effective collaborative learning requires a high intrinsic cognitive load that cannot be

tackled easily by individuals. Students who learned in groups gained a benefit by sharing the high working memory load created by the complex tasks with other group members.

Hypothesis 3: Studying worked examples would be more advantageous than conventional problem solving. This hypothesis flows from cognitive load theory and the worked example effect. It was tested in Experiment 2.

Experiment 1

This experiment tested the hypotheses by investigating the influence of problem complexity on individual and collaborative learning using a worked example strategy. Two types of algebraic problems were created with low or high levels of complexity. Problem complexity was categorized by the number of steps required to complete the solution, and the level of conceptual knowledge required. The topic, solving linear equations, was selected from the National Curriculum of Indonesia as the experiment was conducted in Indonesian schools.

With problems of differing complexity it was feasible that the order in which they were presented could influence subsequent learning. Hence, the problem sequence was counterbalanced in this experiment to avoid sequential learning effects. All participants received a set of worked examples to test Hypotheses 1 and 2. A 2 (learner grouping context: Collaborative vs. Individual) × 2 (level of complexity: Low vs. High) × 2 (task sequence: Low–High complexity vs. High–Low complexity) mixed experimental design was used with level of complexity the repeated measure.

Method

Participants. One hundred eighty-two students from six Year 7 mathematics classes in an Indonesian school in Magetan, East Java, participated in the study. The school followed the national curriculum, and the topics used in the experiment were mandated by the curriculum. The Indonesian national curriculum requests teachers not to use teacher-centered learning methods such as lectures but to use student-centered learning methods such as small group discussions (BNSP, 2006; Depdiknas, 2004; National Ministry of Education, 2006). The participating school indicated that the students were used to studying in small groups in all subjects with varied methods of instruction. The school also indicated that they had allocated students to the six mathematics classes randomly at the beginning of the school year. A team of three mathematics teachers taught specific topics to all six classes, indicating that all students received mathematics instruction from each teacher on set topic blocks throughout the school year.

At the beginning of the school year students were assigned to small learning groups by the mathematics teachers based on having the same gender, and of mixed ability (heterogeneous groupings). Grouping students together according to gender was part of the school's policy for students this age, as it was assumed that boys and girls interact minimally and form single-sex friendships. As friendship groupings can have positive effects on collaboration (see Hanham & McCormick, 2009), it was thus assumed that the group members had developed some level of cohesiveness and familiarity with each other, and could work collaboratively.

These preexisting groups that had been created 3 months earlier by the school, independent of this study, formed the basis for creating the two grouping treatments. Each group was assigned at random to either stay as a group or become uncoupled to study individually. This process produced 79 individual learners and 27 collaborative groups (22 groups of 4, 5 groups of 3, $n = 103$). Both groups and individual learners were then randomly assigned to a specific task sequence of either low–high or high–low complexity. Due to absenteeism 168 students (88 girls, 80 boys) actually participated, with an average age of 12.6 years ($SD = 0.46$). In the low–high complexity sequence, 38 students completed the task individually and 45 students completed the task collaboratively (9 groups of 4, 3 groups of 3). In the high–low complexity sequence, there were 33 students in an individual and 52 students in a collaborative context (10 groups of 4, 4 groups of 3).

Materials. Two types of algebra problems were created based on solving linear equations with differing levels of complexity. Both task types required students to solve a linear equation. The low complexity problem was presented in algebraic notation, but the high complexity problem required an equation to be derived, as it was presented as a word problem. The requirement to translate the words into equations increased complexity.

An example of a low-complexity problem is "*Solve $3n + 10 = 85$, for n.*" An example of an equivalent high-complexity problem is "*Three times the number of Dina's marbles when added to 10 equals eighty-five. How many marbles does Dina have?*" The high-complexity problem required more solution steps, as not only does the equation have to be constructed, a conceptually demanding task, but it also has to be solved. Consequently, this word problem was considered higher in element interactivity (Sweller, 2010; Sweller & Chandler, 1994), because several variables have to be considered simultaneously to construct the equation, although the given problem context may describe operators (symbols) in the constructed equation more meaningfully. In contrast, the low complexity problem does not have this additional task; hence, the algebra rules can be applied in a straightforward fashion. The students in this study had some previous experience with linear equation solving and word problems, but mostly with fewer variables, and not with a combination of constructing and solving equations. For each problem type, instructional and testing materials were constructed.

An instructional materials booklet was designed using a worked example approach. The worked example material used problem pairs, consisting of a worked example and a similar problem to be solved (see Sweller & Cooper, 1985; Trafton & Reiser, 1993). The worked example provided a problem statement and a step-by-step solution to the problem (i.e., algorithm, explanation, final answer) and was written on the left side of the page. The paired problem to be solved was positioned on the right side of the page and consisted of the problem statement only. Final answers for these problems, but not step-by-step solutions, were provided on the same page of the booklet to allow students to know whether they had correctly solved the problem, providing some support consistent with previous research (see Cooper & Sweller, 1987). The relevant instruction was provided directly above each problem. All instructions were in the students' native Indonesian. Appendixes A and B show examples, translated into English, of the format of the low-complexity and the high-complexity worked examples respectively.

The learning material of low-complexity problems consisted of four worked example problem pairs. Hence. the worked example

condition required learners to study 4 worked examples and solve 4 problems overall, whereas the problem solving condition required all 8 problems to be solved. The similar and transfer tests required 4 and 3 problems to be solved, respectively. The internal consistency of the similar test using Cronbach's alpha was .84, and .75 for the transfer test. The transfer test problems consisted of modified equations requiring more solution steps than the similar test problems. The learning material of high-complexity problems consisted of 3 worked example problem pairs. Hence, the worked example condition required learners to study 3 worked examples and solve 3 problems overall, whereas the problem solving condition required all 6 problems to be solved. The similar and transfer tests consisted of 3 and 2 problems, respectively. Cronbach's alpha was .86 for the similar test, and .71 for the transfer test. The transfer test problems had the additional requirement of calculating a subgoal before the goal could be calculated.

To measure cognitive load during acquisition, a self-rating scale of difficulty was used based on the scale developed by Paas (see Paas, 1992; van Gog & Paas, 2008). Furthermore, consistent with recent research, which suggested that multiple recordings produce the most consistent results (see van Gog, Kirschner, Kester, & Paas, 2012), every page of the instructional material had a subjective rating question, written on the bottom line of the page, that asked, "How easy or difficult was it to study and solve these problems? Circle your answer on a scale from 1 = *Extremely easy* to 9 = *Extremely difficult.*" The cognitive load measures collected on each page were added and then averaged to describe the overall student's cognitive load experience in this phase.

Procedure. Before the experimental stage started, all students underwent a preparation period. This initial session was conducted by one of the researchers who was a native Indonesian mathematics teacher. First, students practiced translating a word problem containing one operator into an equation, based on the statement "Bob has 3 more marbles than Wina." The purpose was to activate students' prior knowledge about translating a simple sentence containing a variable into an equation along with the basic algebra rules. The researcher used explicit instruction to explain how to solve this problem.

Second, to familiarize students with instruction using worked examples, four pairs of worked examples, using the same format as in the main experiment, were provided. Each pair consisted of an example to study followed by a similar problem to solve and dealt with translating a simple sentence into an equation. This practice lasted 15 min and then the results of the constructed formula were discussed with the teacher. Immediately afterward, three worked-example pairs for solving simple linear equations by applying one algebra rule were given (e.g., solve $a + 20 = 65$). The results of this 15-min practice were then also discussed with the teacher. This discussion was based on student questions with the researcher responding to the questions without further elaboration.

To complete the preparation period, the teacher then provided an example of a word problem (complex problem), similar to the first problem of the high complexity problem in the learning material. This problem was written on the blackboard. and students were shown how to translate the word problem into a linear equation. The teacher explained that two or more steps were required to transform the linear equation in such a way that it could be solved. However, the step-by-step solution and the final answer were not shown. The whole preparation period was repeated for each class (6 times) by the same researcher.

In the first stage of the experiment (Stage I), students in the Low–High sequence were presented the low-complexity materials first, whereas those in the High–Low sequence were presented the high-complexity materials first. This stage consisted of three phases: acquisition, similar test, and transfer test, which were completed without pauses between them.

Students in each class were separated into two classrooms according to their grouping classifications to begin the acquisition phase, with each classroom supervised by both a teacher from the school and the researcher. First, each student received a worked example booklet specific to their learning condition. Twenty min were allocated for all groups completing low-complexity problems, and 30 min were allocated for high-complexity problems. Before learning commenced, the supervising teacher explained the rules for studying individually or collaboratively, reading from a common script for each strategy.

For individual study, students were told to put an effort into understanding the learning material individually and were not permitted to ask any questions of the other students or the teacher during learning. For collaborative study, students were told by the teacher to discuss the learning material together by reading the task together, eliciting understanding, helping each other, and making sure every member understood the learning material. They were not permitted to ask any questions of other group members or the teacher during learning. For both groups it was also explained how students should complete the cognitive load measures that would appear on each page of their booklet. No feedback was provided during or after the acquisition phase.

Directly following the acquisition phase, the similar and transfer tests were completed individually. All students were given the maximum time period and did not receive any feedback. Fifteen min and 20 min were given to complete the low-complexity similar test and transfer tests, respectively. To complete the high-complexity similar and transfer tests, 20 min were given for each test. After the transfer test, students were given a 15-min break.

Stage II was completed directly after the break, and students switched to the alternate complexity level materials. If students had initially completed the low complexity problems, they then completed those with high complexity next, and vice versa. Allocated times depended on the material and activity as described in Stage I.

During the acquisition phase, group answers were allowed for some groups, and therefore these data were not analyzed, as individual responses were not available for all participants. Scoring for the similar and transfer tests used the following guidelines: For a low-complexity problem, each successful answer had to complete two steps showing two algebraic manipulations. If the answer was entirely correct, a score of 2 was given. If only one step (one strategy) was correctly applied, a score of 1 was given. If the answer did not show any algorithmic validity, a score of 0 was given. For a high-complexity problem, each correct answer had to include three steps. The first was creating the linear equation, while the second and third steps were solving the equation. If the answer was entirely correct, a score of 3 was given. If the equation was correctly created (the first step correct) but only partially solved (1 correct step), a score of 2 was given. If the equation was correctly created (the first step correct) but incor-

rectly solved (0 correct steps), a score of 1 was given. If the equation was incorrectly created but was solved correctly, a score of 1 was given. If the answer did not demonstrate any correct steps, a score of 0 was given. To enable a comparison to be made between the two types of problems, the total scores for each measure were converted into a proportion.

Results and Discussion

A 2 (Collaborative vs. Individual) × 2 (Low–High vs. High–Low complexity sequence) × 2 (Low- vs. High-complexity) ANOVA with repeated measures on the last variable was used to analyze the data. The means (and standard deviations) of test performance and cognitive load ratings are summarized in Table 1.

Cognitive load during acquisition. A significant complexity effect was found, $F(1, 164) = 41.80$, $MSE = 1.40$, $p < .001$, $\eta_p^2 = .203$. The low-complexity materials in the acquisition phase were rated significantly easier ($M = 3.05$, $SD = 1.63$) than the high-complexity materials ($M = 3.89$, $SD = 1.89$). However, no significant main effect was found for learner grouping or task sequence (for both, $F < 1$, ns.). Nor was there a significant interaction between learner grouping and task sequence, nor between learner grouping and problem complexity (for both, $F < 1$, ns.).

A significant interaction effect between problem complexity and task sequence was found, $F(1, 164) = 18.11$, $MSE = 1.40$, $p < .001$, $\eta_p^2 = .099$. Simple effect tests indicated that students reported a higher increase in cognitive load for high-complexity problems compared to low-complexity problems when the task sequence was Low–High, $F(1, 82) = 61.80$, $MSE = 1.38$, $p < .001$, $\eta_p^2 = 0.43$, compared to when the sequence was High–Low, $F(1, 84) = 1.958$, $MSE = 1.45$, $p = .165$, $\eta_p^2 = .023$. As inspection of the means indicates, the higher increase in cognitive load for the higher-complexity problems under a Low-High sequence is primarily due to the relatively low load imposed by low complexity problems when they are presented first.

Similar test results. There was a main effect for task complexity, $F(1, 164) = 50.76$, $MSE = 0.06$, $p < .001$ $\eta_p^2 = .236$. Students scored significantly higher on the low-complexity problems ($M = 0.63$, $SD = 0.34$) than the high-complexity problems ($M = 0.44$, $SD = 0.35$). There was no significant main effect for learner grouping context ($F < 1$, ns.) or task sequence ($F < 1$, ns.). There was no interaction between learner grouping and complexity, $F(1, 164) = 2.18$, $MSE = 0.06$, $p = .142$, $\eta_p^2 = .013$, and all other interaction measures were non-significant (all $F < 1$, ns.).

Transfer test results. A main effect of complexity was found, $F(1, 164) = 13.58$, $MSE = 0.07$, $p < .001$, $\eta_p^2 = .076$. The scores for the low-complexity transfer test ($M = 0.44$, $SD = 0.37$) were significantly greater than those for the high-complexity transfer test ($M = 0.34$, $SD = 0.35$). A learner grouping effect was also found, $F(1, 164) = 7.54$, $MSE = 0.19$, $p = .007$, $\eta_p^2 = .044$. Learning individually ($M = 0.46$, $SD = 0.38$) resulted in better transfer results than learning collaboratively ($M = 0.32$, $SD = 0.34$). A significant interaction between problem complexity and learner grouping context was also found, $F(1, 164) = 19.51$, $MSE = 0.07$, $p < .001$, $\eta_p^2 = .106$ (see Figure 1). Simple effects tests indicated that learning individually resulted in better performance in high-complexity tasks than collaborative learning, $F(1, 166) = 20.59$, $MSE = 0.13$, $p < .001$, $\eta_p^2 = .11$. However, no difference was found for low-complexity tasks ($F < 1$, ns.).

A nonsignificant difference between task sequences was found, $F(1, 164) = 3.57$, $MSE = 0.19$, $p = .06$, $\eta_p^2 = .021$, although the High–Low sequence generating higher scores ($M = 0.41$, $SD = 0.31$) than the Low-High sequence ($M = 0.34$, $SD = 0.32$). A significant interaction effect between the learner grouping context and task sequence was found, $F(1, 164) = 3.89$, $MSE = 0.19$, $p = .05$, partial $\eta^2 = 0.02$. The simple effects test results indicated that individual learning resulted in a better performance than group learning, when the learning sequence was High–Low, $F(1, 83) = 12.35$, $MSE = 0.17$, $p = .001$, $\eta_p^2 = .13$. When the task sequence was Low–High, no significant differences were found, ($F < 1$, ns.). Moreover, a significant interaction between the task complexity and task sequence was found, $F(1, 164) = 4.15$, $MSE = 0.07$, $p = .043$, $\eta_p^2 = .025$. The simple effects test results indicated a significant difference of low and high complexity transfer performance when the learning sequence was Low–High, $F(1, 82) = 16.20$, $MSE = 0.08$, $p = .001$, $\eta_p^2 = .17$. When the task sequence was High–Low, no significant differences were found, $F(1, 84) = 3.48$, $MSE = 0.07$, $p = .07$, $\eta_p^2 = .04$).

An important aim of this experiment was to create two types of tasks based on simultaneous equations that had two levels of complexity. The results indicated that this aim was supported as students scored significantly higher on the low-complexity problems compared to high-complexity problems on both tests. Additionally, students also experienced a considerably lower cognitive load when learning using low-complexity problems compared to high-complexity problems. Therefore, it is likely that the higher-complexity task, with more steps for the solution, has a higher level of element interactivity.

The first hypothesis of this experiment (Hypothesis 1) predicted that students would benefit from studying worked examples collaboratively rather than individually. No evidence was found during testing to support this prediction. In contrast, a number of

Table 1
Means (and Standard Deviations) for Test Results and Cognitive Load Ratings in Experiment 1

Condition	Cognitive load (1–9)		Similar test (0–1)		Transfer test (0–1)	
	Low-complexity	High-complexity	Low-complexity	High-complexity	Low-complexity	High-complexity
Low–High sequence						
Collaborative	2.71 (1.44)	4.49 (2.26)	.64 (.35)	.42 (.32)	.47 (.41)	.17 (.33)
Individual	2.89 (1.56)	3.92 (1.87)	.57 (.36)	.44 (.35)	.38 (.38)	.34 (.38)
High–Low sequence						
Collaborative	3.31 (1.67)	3.46 (1.60)	.65 (.30)	.41 (.34)	.41 (.35)	.23 (.28)
Individual	3.30 (1.85)	3.73 (1.63)	.64 (.33)	.47 (.38)	.51 (.35)	.59 (.41)

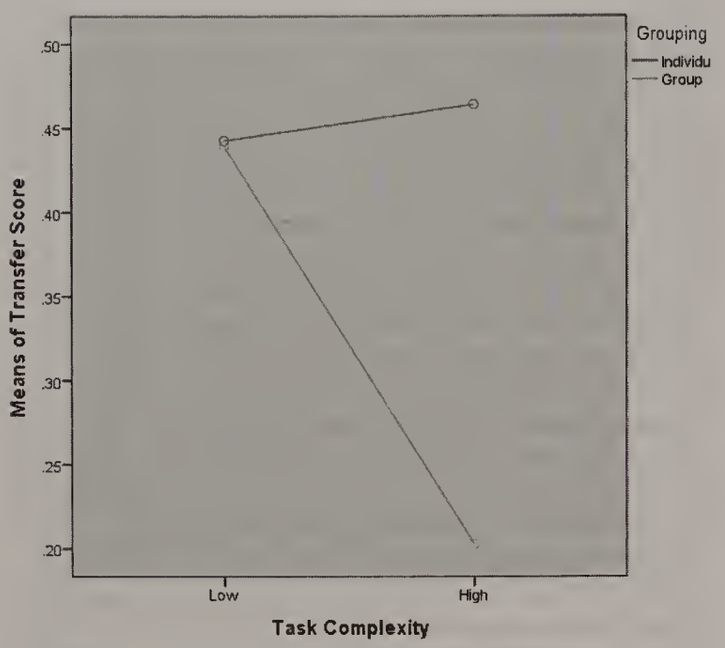


Figure 1. Interaction between task complexity and grouping context on transfer test scores in Experiment 1.

results indicated a reverse effect. On the transfer test, students who studied individually performed significantly higher than those who studied collaboratively, and more specifically on the higher-complexity tasks. Furthermore, for the High–Low order of study, individual study resulted in a significant advantage.

Hypothesis 2 predicted that the effectiveness of collaborative learning would be increased by task complexity. No support for this hypothesis was found, as collaborative learning was not found to be superior to individual learning on any specific task. The significant interaction effect on transfer problems indicated that individual study was more advantageous than collaborative study, for high-complexity problems—the reverse of what was expected. It was concluded, based on these results in the current context, that when using worked examples, collaborative study was a disadvantage.

Experiment 2

The results of Experiment 1 suggested that worked examples may not be enhanced by collaborative learning, but it was notable that worked examples were not compared with a problem solving control group as is usually the case in worked-examples research. Because we did not test for the worked example effect in Experiment 1, it is possible that the worked example approach was unsuitable for this match of topic and learner, leading to no learning advantage for using worked examples. To rule out this possibility, Experiment 2 included a problem-solving treatment. This enabled the first two hypotheses to be tested again using a 2 (instructional strategy: Worked Example vs. Problem Solving) × 2 (grouping contexts: Collaborative vs. Individual) × 2 (level of complexity: Low vs. High) mixed experimental design. Because in Experiment 1 few differences were found by balancing the sequencing of problem types, only the complex-simple sequence (High–Low complexity task sequence) was used in Experiment 2, as it produced the most significant interactions.

With the introduction of a problem solving treatment it was possible to test for a worked example effect (see Atkinson et al., 2000; P. Kirschner et al., 2006; Renkl, 2014a, 2014b). In other words, it was predicted that worked examples would be advantageous compared to conventional problem solving (Hypothesis 3).

Method

Participants. One hundred twenty-two students from four Year 7 classrooms in an Indonesian school, in Kudus, Central Java, participated in the study. Consistent with the sample in the previous experiment, the school had a similar organization and followed the same national curriculum. Consistent with Experiment 1, a team of three mathematics teachers taught all classes according to set topics. Also similar to the participants in Experiment 1, collaborative learning was reported as a common learning strategy, not only in mathematics classes but also in other subjects. As was the case in Experiment 1, the small groups were created at the beginning of the school year independently of this experiment, and composed of mixed ability students with the same gender. Students were assumed to be familiar with each other since they had been in the same groupings for more than five months.

First, students were randomly allocated into individual (n = 63) or collaborative (n = 60, 9 groups of 4, and 8 groups of 3) learning conditions, and then randomly assigned into worked example or problem solving groups. Five students were excluded from the analysis because they did not complete all experimental stages, leaving 118 students (46 girls, 72 boys) with an average age of 12.50 years (SD = 0.55). Thirty students studied worked examples individually, 29 solved problems individually, 31 (4 groups of 4, and 5 groups of 3) studied worked examples collaboratively, and 28 (5 groups of 4, 2 groups of 3, and 1 group of 2) solved problems collaboratively.

Learning materials and procedure. The materials used in this experiment were identical to Experiment 1, except that a new conventional problem-solving group was introduced. Problem-solving acquisition booklets for both levels of complexity were designed, based on the worked examples booklets. Where in Experiment 1 for each problem pair, the first problem had a fully worked example given, this solution was no longer provided. Instead, this problem now had to be solved by participants during the acquisition phase. Hence, for the worked example condition, students studied a problem, and solved a similar problem; for the problem solving condition, both problems had to be solved without solutions being shown. Each problem pair was placed on a single page, positioned identically to the worked example material except that no solutions were shown. Students were instructed to solve each problem. Equivalent to the worked example booklet, the final answer of every problem was provided on each page.

The similar test and the transfer test materials were identical to the low- and high-complexity problems used in Experiment 1, and the allocated times remained the same. The internal consistency of the tests was measured again using Cronbach's alpha for this sample. For the low-complexity problems, the values were .88 for the similar test and .80 for the transfer test. For the high-complexity problems, the values were .82 for the similar test and .67 for the transfer test.

The procedures used in this experiment were identical to Experiment 1. The only difference, apart from introducing additional

problem-solving groups, was that only the complex-simple sequence (High–Low complexity task sequence) was used for the two types of problems, as this sequence previously produced significant interactions.

Results and Discussion

A 2 (Worked Example vs. Problem Solving) × 2 (Collaborative vs. Individual) × 2 (Low- vs. High-complexity) ANOVA with repeated measures on the last variable was used to analyze the data. The means (and standard deviations) of test performance and cognitive load ratings are summarized in Table 2.

Cognitive load during acquisition results. A main effect of instructional strategy was obtained, $F(1, 114) = 90.96$, $MSE = 3.82$, $p < .001$, $\eta_p^2 = .44$, where the worked example conditions ($M = 3.80$, $SD = 1.56$) generated significantly lower cognitive load scores (difficulty scale) than the problem solving conditions ($M = 6.23$, $SD = 2.08$). No significant effect for learner grouping contexts was found, $F(1, 114) = 1.97$, $MSE = 3.82$, $p = .164$, $\eta_p^2 = .02$. The low-complexity problems ($M = 4.47$, $SD = 1.93$) generated significantly less cognitive load than high-complexity problems ($M = 5.56$, $SD = 1.71$), $F(1, 114) = 23.5$, $MSE = 2.99$, $p < .001$, $\eta_p^2 = .17$.

A significant 3-way interaction was found, $F(1, 114) = 3.99$, $MSE = 2.99$, $p = .048$, $\eta_p^2 = .034$. Simple effects tests showed that individual learners experienced a significantly higher cognitive load than learners in the collaborative context when learning the high-complexity problems using worked examples, $F(1, 59) = 4.34$, $MSE = 2.45$, $p = .041$, $\eta_p^2 = .069$ (see Figure 2), but no other significant effects were found.

Similar test results. A significant worked example effect was found, $F(1, 114) = 24.93$, $MSE = 0.111$, $p < .001$, $\eta_p^2 = .18$, as studying worked examples ($M = 0.53$, $SD = 0.27$) was found to be superior to problem solving ($M = 0.32$, $SD = 0.28$). No significant effect for the learner grouping context was found ($F < 1$, ns.). However, there was an interaction effect between instructional strategy and learner grouping context, $F(1, 114) = 5.92$, $MSE = 0.111$, $p = .017$, $\eta_p^2 = .049$ (see Figure 3). Simple effects tests revealed that there were no significant differences between the learner grouping contexts when students studied worked examples, $F(1, 59) = 2.56$, $MSE = 0.10$, $p = .115$, $\eta_p^2 = .042$. There was nonsignificant difference between means, with a small to medium effect size, $F(1, 55) = 3.32$, $MSE = 0.12$, $p = .07$, $\eta_p^2 = .057$, in favor of collaborative learning when students studied through the problem-solving format. That difference can be assumed to have been the primary cause of the significant interaction.

A significant effect of complexity was found, $F(1, 114) = 23.72$, $MSE = 0.04$, $p < .001$, $\eta_p^2 = .172$. Students performed significantly higher in low-complexity problems ($M = 0.49$, $SD = 0.29$) than in high-complexity problems ($M = 0.36$, $SD = 0.25$).

A 3-way interaction effect was also found, $F(1, 114) = 8.83$, $MSE = 0.04$, $p = .004$, $\eta_p^2 = .072$, caused by the significant differences found in low-complexity tests (see Figure 4). For worked examples, individual study was superior to collaborative study, $F(1, 59) = 7.01$, $MSE = 0.08$, $p = .01$, $\eta_p^2 = .106$, replicating the results of Experiment 1, but when problem solving, collaborative study was superior to individual study, $F(1, 55) = 4.67$, $MSE = 0.09$, $p = .035$, $\eta_p^2 = .078$. No significant differences were found for high-complexity problems ($F < 1$, ns., for both).

Transfer test results. There was no worked example effect, $F(1, 114) = 2.06$, $MSE = 0.104$, $p = .154$, $\eta_p^2 = .018$, nor a learner grouping context effect ($F < 1$, ns.) However, there was a significant interaction between the instructional strategy and the learner grouping context, $F(1, 114) = 8.60$, $MSE = 0.104$, $p = .004$, $\eta_p^2 = .070$ (see Figure 5). The simple effects test indicated a significant difference for worked examples, $F(1, 59) = 4.81$, $MSE = 0.08$, $p = .032$, $\eta_p^2 = .075$, where individual study again was superior to collaborative study. For problem solving, a significant effect again was found in favor of collaborative learning, $F(1, 55) = 3.96$, $MSE = 0.13$, $p = .05$, $\eta_p^2 = .067$.

A main effect of complexity was found, $F(1, 114) = 33.64$, $MSE = 0.03$, $p < .001$, $\eta_p^2 = .23$. Students performed significantly higher in the low-complexity transfer problems ($M = 0.31$, $SD = 0.27$) than in the high-complexity problems ($M = 0.18$, $SD = 0.23$). A significant interaction effect between the instructional strategy and problem complexity was also found, $F(1, 114) = 9.75$, $MSE = 0.03$, $p = .002$, $\eta_p^2 = .08$ (see Figure 6). The simple effects tests indicated that for the high-complexity transfer problems, worked examples led to a significantly higher performance than problem solving, $F(1, 116) = 7.96$, $MSE = 0.06$, $p = .006$, $\eta_p^2 = .064$, but for low-complexity transfer problems, there were no significant differences between the learning strategies ($F < 1$, ns.).

No overall support was found for Hypothesis 1 that students would benefit from studying collaboratively rather than individually when using worked examples. Instead, the reverse result was obtained, with individual study superior to collaborative study on both similar and transfer tests for high complexity problems. Interestingly, this superiority was associated with a higher cognitive load for individual study. Normally, a lower cognitive load is associated with improved performance. Future work will be re-

Table 2
Means (and Standard Deviations) for Test Results and Cognitive Load Ratings in Experiment 2

	Cognitive load (1–9)		Similar test (0–1)		Transfer test (0–1)	
Condition	Low-complexity	High-complexity	Low-complexity	High-complexity	Low-complexity	High-complexity
Worked examples						
Collaborative	3.16 (1.32)	4.06 (1.34)	.49 (.27)	.48 (.24)	.24 (.27)	.19 (.20)
Individual	3.07 (1.80)	4.90 (1.77)	.68 (.30)	.47 (.27)	.36 (.22)	.30 (.23)
Problem solving						
Collaborative	5.43(1.99)	6.68 (1.85)	.55 (.31)	.31 (.30)	.40 (.30)	.16 (.35)
Individual	6.21(2.60)	6.59 (1.86)	.31 (.32)	.16 (.23)	.22 (.29)	.07 (.13)

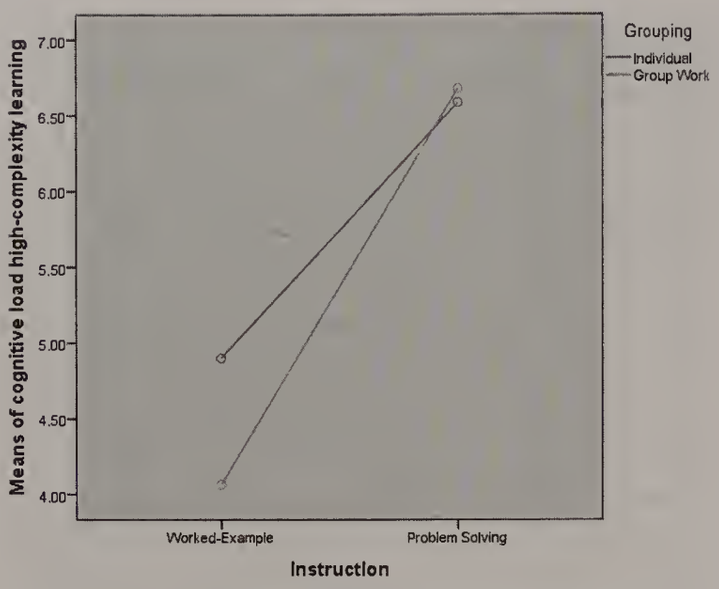


Figure 2. Interaction between instructional strategy and grouping context on cognitive load during learning high complexity material in the acquisition phase in Experiment 2.

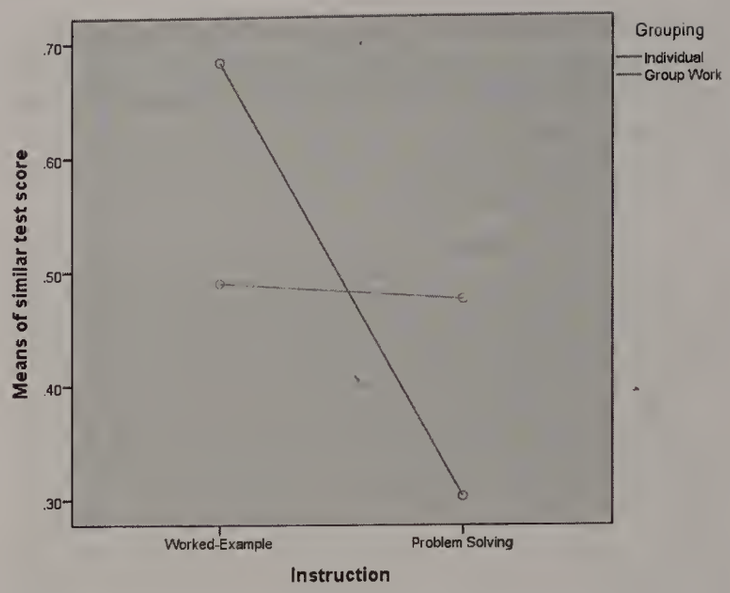


Figure 4. Interaction between instructional strategy and grouping context for similar test scores on low-complexity task in Experiment 2.

quired to establish whether this result is replicable. Nevertheless, based on interaction effects, there was some advantage for collaboration. Collaborative learning was superior to individual learning on similar and transfer tests when students initially used the problem solving strategy.

There was no support for Hypothesis 2 that the effectiveness of collaborative learning would be increased by task complexity. It was found that for the problem solving strategy, collaborative learners performed better than individual learners on similar tests for the low-complexity problems. For the worked example strategy, however, individual learners performed better than collaborative learners for low-complexity similar tests. No effects were obtained using high-complexity problems.

Hypothesis 3 predicted that students would benefit from studying worked examples rather than problem solving. An overall main effect was found in support of this hypothesis for the similar test problems. Although no main effect for worked examples was found on the transfer test phase, an interaction effect indicated that on high-complexity transfer problems, those who studied worked examples scored higher than those who initially solved problems. Furthermore, during the similar test phase, the cognitive load was found to be lower when studying worked examples rather than solving problems.

In summary, this experiment confirmed that the worked example strategy was superior to the problem solving strategy. When the material was higher in complexity, learning by worked examples in an individual setting was advantageous compared to col-

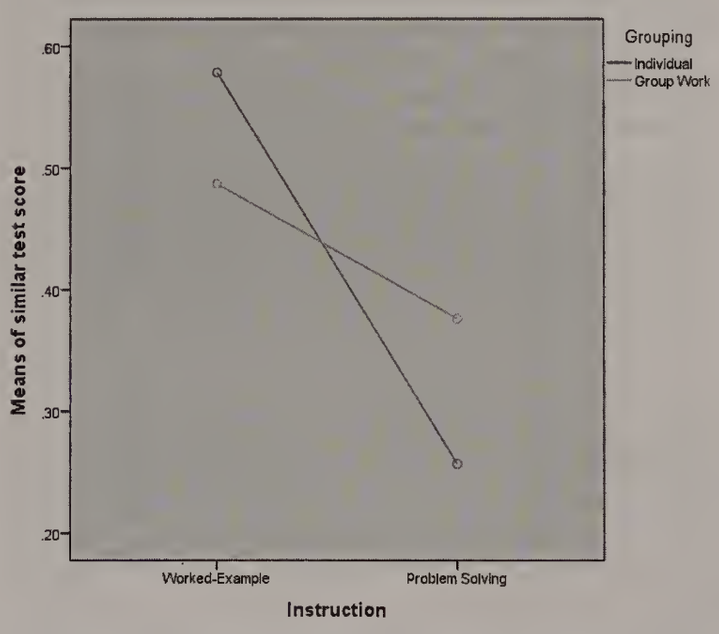


Figure 3. Interaction between instructional strategy and grouping context on similar test scores in Experiment 2.

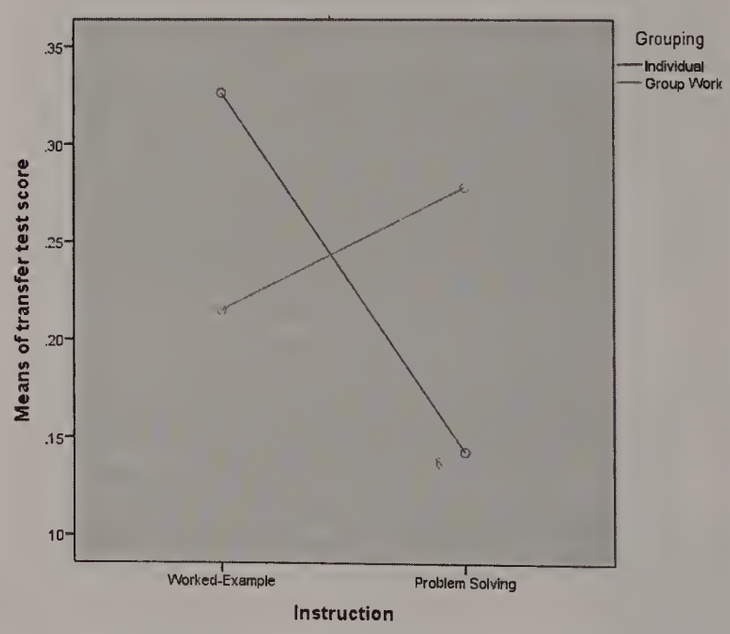


Figure 5. Interaction between instructional strategy and grouping context on transfer test scores in Experiment 2.

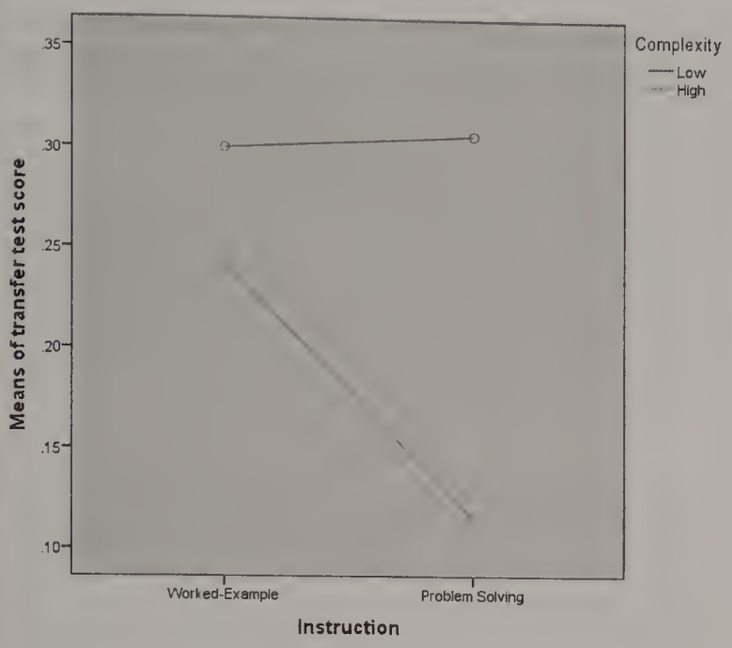


Figure 6. Interaction between instructional strategy and task complexity on transfer test scores in Experiment 2.

laborative learning. When the material was lower in complexity, some benefits were found for problem solving in groups.

General Discussion

Summary of Evidence in Support of the Hypotheses

Three hypotheses were tested over the two experiments, and the overall evidence is summarized below.

The first hypothesis predicted that *worked examples will be enhanced by studying collaboratively compared to studying individually.* This hypothesis was examined in both experiments and was not supported. When all students were tested individually on the similar and transfer tests, no superiority was found. In contrast, the results of Experiment 1 (transfer test performance in a worked example environment) showed that individual learners had a significant advantage over collaborative learners, while Experiment 2 indicated an advantage for individual study on both similar and transfer tasks for high complexity problems. Consequently, Hypothesis 1 was rejected.

The second hypothesis predicted that *the effectiveness of collaborative learning is increased by task complexity.* This hypothesis was examined in Experiments 1 and 2. As reported, these experiments used two tasks with different levels of complexity. It was predicted that there would be interactions between task complexity and the effectiveness of collaborative learning. Although interactions were found, follow-up tests indicated that collaborative learning was not superior to individual learning on the more complex tasks. In fact, in Experiment 1 (transfer test performance), evidence emerged that for the more complex tasks individual learning was superior to collaborative learning. Thus, Hypothesis 2 was rejected.

The final hypothesis predicted that *studying worked examples would be more advantageous than conventional problem solving.* This hypothesis was examined in Experiment 2. The results indicated evidence in support of the effectiveness of a worked example strategy compared to a problem solving strategy (the worked example effect), as predicted. Worked examples were found to be more effective during the similar test phase. Furthermore, students using the worked example strategy experienced a lower cognitive load for the similar test phase. A significant interaction was also found. Students who originally studied worked examples had higher scores on the high complexity transfer problems than those who originally were asked to solve problems.

As summarized above, no evidence was found that when using worked examples, collaborative learning was significantly superior to individual learning. In contrast, some evidence emerged that individual contexts were superior. A perspective from evolutionary educational psychology (Geary, 1995, 2002) can be used to explain why collaborative learning was rarely superior to individual learning. Geary argues that in social interactions, students develop their biologically primary knowledge rather than the assigned biologically secondary knowledge. In other words, they become more adept at their social interaction, which is an evolutionary primary skill, rather than the assigned mathematics task, which is an evolutionary secondary skill, and requires considerable conscious effort to learn. It was expected that group interactions would have generated superior sense making and reorganization of the information provided. However, there was no evidence for this suggestion. It is possible that worked examples provide sufficient information, rendering collaboration unnecessary.

The study also examined if the effectiveness of collaborative learning was important when dealing with complex tasks. High-complexity problems were argued to increase active social interaction during collaborative learning. As Hypothesis 2 was rejected, it can be concluded that for the complexity levels used, neither low- nor high-complexity tasks improved collaborative learning compared with individual learning. In fact, it was found on several occasions that for high-complexity tasks, individual learning led to higher performance than collaborative learning.

Moreover, it was found that collaborative learning only had an advantage over individual learning during problem solving. Collaborative learners scored higher than individual learners after having acquired their initial knowledge through problem solving. It is also notable that the collaborative advantage occurred only on low-complexity materials. It is possible that the low-complexity problem solving imposed a lower cognitive load and thus could be managed in a collaborative learning setting.

The study also tested for a worked example effect. The evidence obtained in this study is consistent with cognitive load theory research, demonstrating that overall the worked example strategy was superior to a problem solving strategy. Worked examples in general can be used in individual or collaborative learning contexts, replicating a previous finding (Retnowati et al., 2010). It is important to note, however, that the various interactions identified in this study indicated that the worked example strategy was best used in individual rather than collaborative settings, particularly for high-complexity problems.

Collaborative learning creates conditions where students in a group are expected to discuss the learning material, which can be done by giving/receiving elaborated explanations (Cohen, 1994; Webb, 1991, 2009). However, worked examples contain step-by-step explanations to reach a problem solution, so discussing worked examples may have a redundant element (Chandler &

Sweller, 1991). Worked examples are unnecessary if members of the group can "borrow" (using the cognitive load theory borrowing and reorganizing principle) the information required to learn or solve the given problem from the other group members. Similarly, group interactions may well help enhance the reorganization of new information.

While the current results were theoretically coherent and largely consistent, they will require replication in different contexts using different populations and materials. We have established that at least under some circumstances, collaboration when studying worked examples has negative rather than positive effects, while collaboration using problem solving can have positive effects. As far as we are aware, this finding is novel. We have interpreted these findings in terms of redundancy (Nihalani et al., 2011). Learners studying worked examples do not need additional information from collaborators to assist them when studying. Such additional, redundant information may have negative rather than positive effects, leading to an expertise reversal effect (Kalyuga et al., 2003). In contrast, when problem solving in the absence of worked examples, information from collaborators may be beneficial. Whether collaboration when studying worked examples is advantageous under different circumstances requires additional data. For example, exceptionally complex worked examples may benefit from a collaborative approach.

One potential limitation of the study occurred because we wanted to examine an authentic learning environment, and therefore some recommended steps, such as group processing training to prepare for effective collaboration, were not followed (see Johnson & Johnson, 1994). It is feasible that the group processes conducted by this sample were not sufficient to optimize the impact of collaboration. Nevertheless, the groups had been working together in mathematics classes for 3 (Experiment 1) and 5 (Experiment 2) months and so had experience learning in their groups. Furthermore, the finding that collaboration was superior to individual study for the problem-solving strategy suggests that there were benefits, and therefore a certain amount of effective collaborative behavior can be assumed to have been present. To collect additional data on this issue was outside the scope of the present study, but is a topic for further investigation. Furthermore, replicating this study with collaborative groups further prepared according to the steps often recommended for effective collaboration should be also be informative.

Many effective collaboration tasks consist of realistic ill-structured problems (Hmelo-Silver, 2004), have high complexity (F. Kirschner et al., 2009a), or cannot be completed by individuals (Cohen, 1994). In contrast, the tasks chosen for this study (middle school equation solving) did not contain all these characteristics. Nevertheless, we did test for the effect of complexity, and significant differences were found between the two complexity levels. Future studies could include richer problem solving tasks with more ill-defined goals as recommended. Also, delayed tests could be included in future studies to judge the permanence of learning, although it should be noted that worked examples have been found to provide robust learning longevity (see Chen et al., 2016a).

With respect to educational implications, our main research question was Can collaborative learning improve the effectiveness of worked examples? Under the given conditions, the answer is no. Worked examples seem to be most effective in individual settings. Asking learners to discuss worked examples may be redundant because they have already obtained the necessary information from an instructor via the worked example. Regarding problem complexity, individual study seemed to be most appropriate for the complex problems, although collaboration was helpful when problem solving (the inferior strategy), presumably because collaboration permitted learners to obtain missing information from other learners. Hence, collaboration may be advantageous when problem solving because to some extent it is able to provide learners with missing guidance.

In conclusion, there appear to be limits to the conditions under which collaborative learning is effective. Those limits should be considered when encouraging learners to study collaboratively.

References

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research, 70,* 181–214. http://dx.doi.org/10.3102/00346543070002181

Ayres, P., & Sweller, J. (2013). The worked example effect. In J. A. C. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 408–410). Oxford, England: Routledge.

Barron, B. (2000). Achieving coordination in collaborative problem-solving groups. *Journal of the Learning Sciences, 9,* 403–436. http://dx.doi.org/10.1207/S15327809JLS0904_2

BNSP. (2006). *Panduan penyusunan kurikulum tingkat satuan pendidikan jenjang pendidikan dasar dan menengah* [Guideline for developing curriculum for elementary and middle education]. Jakarta, Indonesia: Badan Nasional Standar Pendidikan.

Carroll, W. M. (1994). Using worked examples as an instructional support in the algebra classroom. *Journal of Educational Psychology, 86,* 360–367. http://dx.doi.org/10.1037/0022-0663.86.3.360

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8,* 293–332. http://dx.doi.org/10.1207/s1532690xci0804_2

Chen, O., Kalyuga, S., & Sweller, J. (2015). The worked example effect, the generation effect, and element interactivity. *Journal of Educational Psychology, 107,* 689–704. http://dx.doi.org/10.1037/edu0000018

Chen, O., Kalyuga, S., & Sweller, J. (2016a). Relations between the worked example and generation effects on immediate and delayed tests. *Learning and Instruction, 45,* 20–30. http://dx.doi.org/10.1016/j.learninstruc.2016.06.007

Chen, O., Kalyuga, S., & Sweller, J. (2016b). The expertise reversal effect is a variant of the more general element interactivity effect. *Educational Psychology Review.* Advance online publication. http://dx.doi.org/10.1007/s10648-016-9359-1

Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research, 64,* 1–35. http://dx.doi.org/10.3102/00346543064001001

Cooper, G., & Sweller, J. (1987). The effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology, 79,* 347–362. http://dx.doi.org/10.1037/0022-0663.79.4.347

Davidson, N., & Kroll, D. L. (1991). An overview of research on cooperative learning related to mathematics. *Journal for Research in Mathematics Education, 22,* 362–365. http://dx.doi.org/10.2307/749185

De Corte, E. (2004). Mainstreams and perspectives in research on learning (mathematics) from instruction. *Applied Psychology, 53,* 279–310. http://dx.doi.org/10.1111/j.1464-0597.2004.00172.x

Depdiknas. (2004). *Kurikulum 2004 untuk sekolah menengah pertama dan madrasah tsanawiyah* [2004 Curriculum for junior high school and Islamic junior high school]. Jakarta, Indonesia: Departemen Pendidikan Nasional.

Geary, D. C. (1995). Reflections of evolution and culture in children's cognition. Implications for mathematical development and instruction.

American Psychologist, 50, 24–37. http://dx.doi.org/10.1037/0003-066X.50.1.24

Geary, D. C. (2002). Principles of evolutionary educational psychology. *Learning and Individual Differences, 12,* 317–345. http://dx.doi.org/10.1016/S1041-6080(02)00046-8

Geary, D. C. (2008). An evolutionarily informed education science. *Educational Psychologist, 43,* 179–195. http://dx.doi.org/10.1080/00461520802392133

Geary, D. (2012). Evolutionary educational psychology. In K. Harris, S. Graham, & T. Urdan (Eds.), *APA Educational Psychology Handbook* (Vol. 1, pp. 597–621). Washington, DC: American Psychological Association.

Gillies, R. M. (2003). Structuring cooperative group work in classrooms. *International Journal of Educational Research, 39,* 35–49. http://dx.doi.org/10.1016/S0883-0355(03)00072-7

Goos, M. (2004). Learning mathematics in a classroom community of inquiry. *Journal for Research in Mathematics Education, 35,* 258–291. http://dx.doi.org/10.2307/30034810

Hanham, J., & McCormick, J. (2009). Group work in schools with close friends and acquaintances: Linking self-processes with group processes. *Learning and Instruction, 19,* 214–227. http://dx.doi.org/10.1016/j.learninstruc.2008.04.002

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review, 16,* 235–266.

Johnson, D. W., & Johnson, R. T. (1994). *Learning together and alone: Cooperative, competitive and individualistic learning.* Boston, MA: Allyn & Bacon.

Johnson, D. W., & Johnson, R. T. (2002). Learning together and alone: Overview and meta-analysis. *Asia Pacific Journal of Education, 22,* 95–105. http://dx.doi.org/10.1080/0218879020220110

Johnson, D. W., Johnson, R. T., & Smith, K. A. (1998). Cooperative learning returns to college: What evidence is there that it works? *Change: The Magazine of Higher Learning, 30,* 26–35. http://dx.doi.org/10.1080/00091389809602629

Johnson, D. W., Maruyama, G., Johnson, R., Nelson, D., & Skon, L. (1981). The effects of cooperative, competitive, and individualistic goal structures on achievement: A meta-analysis. *Psychological Bulletin, 89,* 47–62. http://dx.doi.org/10.1037/0033-2909.89.1.47

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38,* 23–31. http://dx.doi.org/10.1207/S15326985EP3801_4

Kirschner, F., Paas, F., & Kirschner, P. A. (2009a). A cognitive load approach to collaborative learning: United brains for complex tasks. *Educational Psychology Review, 21,* 31–42. http://dx.doi.org/10.1007/s10648-008-9095-2

Kirschner, F., Paas, F., & Kirschner, P. A. (2009b). Individual and group-based learning from complex cognitive tasks: Effects on retention and transfer efficiency. *Computers in Human Behavior, 25,* 306–314. http://dx.doi.org/10.1016/j.chb.2008.12.008

Kirschner, F., Paas, F., & Kirschner, P. A. (2011). Task complexity as a driver for collaborative learning efficiency: The collective working-memory effect. *Applied Cognitive Psychology, 25,* 615–624. http://dx.doi.org/10.1002/acp.1730

Kirschner, F., Paas, F., Kirschner, P. A., & Janssen, J. (2011). Differential effects of problem-solving demands on individual and collaborative learning outcomes. *Learning and Instruction, 21,* 587–599.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41,* 75–86. http://dx.doi.org/10.1207/s15326985ep4102_1

Kreijns, K., Kirschner, P. A., & Jochems, W. (2003). Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: A review of the research. *Computers in Human Behavior, 19,* 335–353. http://dx.doi.org/10.1016/S0747-5632(02)00057-2

Kyun, S., Kalyuga, S., & Sweller, J. (2013). The effect of worked examples when learning to write essays in English literature. *Journal of Experimental Education, 81,* 385–408. http://dx.doi.org/10.1080/00220973.2012.727884

Mayer, R. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 43–71). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139547369.005

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics.* Reston, VA: Author.

National Ministry of Education. (2006). *Peraturan Menteri Pendidikan Nasional Republik Indonesia No. 22 Tahun* 2006 (Regulation of the Minister of National Education of the Republic of Indonesia No. 22 of 2006).

Nihalani, P., Mayrath, M., & Robinson, D. (2011). When feedback harms and collaboration helps in computer simulation environments: An expertise reversal effect. *Journal of Educational Psychology, 103,* 776–785. http://dx.doi.org/10.1037/a0025276

Oksa, A., Kalyuga, S., & Chandler, P. (2010). Expertise reversal effect in using explanatory notes for readers of Shakespearean text. *Instructional Science, 38,* 217–236. http://dx.doi.org/10.1007/s11251-009-9109-6

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive load approach. *Journal of Educational Psychology, 84,* 429–434. http://dx.doi.org/10.1037/0022-0663.84.4.429

Paas, F., & Sweller, J. (2012). An evolutionary upgrade of cognitive load theory: Using the human motor system and collaboration to support the learning of complex cognitive tasks. *Educational Psychology Review, 24,* 27–45. http://dx.doi.org/10.1007/s10648-011-9179-2

Paas, F., & van Merriënboer, J. J. G. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86,* 122–133. http://dx.doi.org/10.1037/0022-0663.86.1.122

Plass, J. L., O'Keefe, P. A., Homer, B. D., Case, J., Hayward, E. O., Stein, M., & Perlin, K. (2013). The impact of individual, competitive, and collaborative mathematics game play on learning, performance, and motivation. *Journal of Educational Psychology, 105,* 1050–1066. http://dx.doi.org/10.1037/a0032688

Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology, 88,* 144–161. http://dx.doi.org/10.1037/0022-0663.88.1.144

Renkl, A. (2014a). Toward an instructionally oriented theory of example-based learning. *Cognitive Science, 38,* 1–37. http://dx.doi.org/10.1111/cogs.12086

Renkl, A. (2014b). The worked examples principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 391–412). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139547369.020

Retnowati, E., Ayres, P., & Sweller, J. (2010). Worked example effects in individual and group work settings. *Educational Psychology, 30,* 349–367. http://dx.doi.org/10.1080/01443411003659960

Rourke, A., & Sweller, J. (2009). The worked-example effect using ill-defined problems: Learning to recognise designers' styles. *Learning and Instruction, 19,* 185–199. http://dx.doi.org/10.1016/j.learninstruc.2008.03.006

Schreiber, L. M., & Valle, B. E. (2013). Social constructivist teaching strategies in the small group classroom. *Small Group Research, 44,* 395–411. http://dx.doi.org/10.1177/1046496413488422

Slavin, R. E. (1995). *Cooperative learning: Theory, research, and practice* (2nd ed.). Boston, MA: Allyn & Bacon.

Staples, M. (2007). Supporting whole-class collaborative inquiry in a secondary mathematics classroom. *Cognition and Instruction, 25,* 161–217. http://dx.doi.org/10.1080/07370000701301125

Sweller, J. (2010). Element interactivity and intrinsic, extraneous and germane cognitive load. *Educational Psychology Review, 22,* 123–138. http://dx.doi.org/10.1007/s10648-010-9128-5

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4419-8126-4

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction, 12,* 185–233. http://dx.doi.org/10.1207/s1532690xci1203_1

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction, 2,* 59–89. http://dx.doi.org/10.1207/s1532690xci0201_3

Sweller, J., & Sweller, S. (2006). Natural information processing systems. *Evolutionary Psychology, 4,* 434–458. http://dx.doi.org/10.1177/147470490600400135

Tarmizi, R. A., & Sweller, J. (1988). Guidance during mathematical problem solving. *Journal of Educational Psychology, 80,* 424–436. http://dx.doi.org/10.1037/0022-0663.80.4.424

Trafton, J. G., & Reiser, B. J. (1993). *The contributions of studying examples and solving problems to skill acquisition*. Paper presented at the Fifteenth Annual Conference of the Cognitive Science Society, Hillsdale, NJ.

Van den Bossche, P., Gijselaers, W. H., Segers, M., & Kirschner, P. A. (2006). Social and cognitive factors driving teamwork in collaborative learning environments: Team learning beliefs and behaviors. *Small Group Research, 37,* 490–521. http://dx.doi.org/10.1177/1046496406292938

van Gog, T., & Kester, L. (2012). A test of the testing effect: Acquiring problem-solving skills from worked examples. *Cognitive Science, 36,* 1532–1541. http://dx.doi.org/10.1111/cogs.12002

van Gog, T., Kirschner, F., Kester, L., & Paas, F. (2012). Timing and frequency of mental effort measurement: Evidence in favour of repeated measures. *Applied Cognitive Psychology, 26,* 833–839. http://dx.doi.org/10.1002/acp.2883

van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist, 43,* 16–26. http://dx.doi.org/10.1080/00461520701756248

Ward, M., & Sweller, J. (1990). Structuring effective worked examples. *Cognition and Instruction, 7,* 1–39. http://dx.doi.org/10.1207/s1532690xci0701_1

Webb, N. M. (1991). Task-related verbal interaction and mathematics learning in small groups. *Journal for Research in Mathematics Education, 22,* 366–389. http://dx.doi.org/10.2307/749186

Webb, N. M. (2009). The teacher's role in promoting collaborative dialogue in the classroom. *British Journal of Educational Psychology, 79,* 1–28. http://dx.doi.org/10.1348/000709908X380772

Weinberger, A., Stegmann, K., & Fischer, F. (2007). Knowledge convergence in collaborative learning: Concept and assessment. *Learning and Instruction, 17,* 416–426. http://dx.doi.org/10.1016/j.learninstruc.2007.03.007

Yackel, E., Cobb, P., & Wood, T. (1991). Small-group interactions as a source of learning opportunities in second-grade mathematics. *Journal for Research in Mathematics Education, 22,* 390–408. http://dx.doi.org/10.2307/749187

Zhang, L., Ayres, P., & Chan, K. (2011). Examining different types of collaborative learning in a complex computer-based environment: A cognitive load approach. *Computers in Human Behavior, 27,* 94–98. http://dx.doi.org/10.1016/j.chb.2010.03.038

Appendix A

Example of the Low-Complexity Learning Material Using the Worked Example Instruction

Study this example	Solve this problem	Final answer
Solve $3p + 10 = 85$ $3p + 10-\mathbf{10} = 85-\mathbf{10}$ [subtract **10** from both sides] $3p + 0 = 75$ $3p = 75$ $\frac{3p}{3} = \frac{75}{3}$ [divide both sides by **3**] $p = 25$ Hence, the solution is $p = 25$.	$4a + 13 = 65$	$a=13$

Appendix B

Example of the High-Complexity Learning Material Using the Worked Example Instruction

Study this example	Solve this problem	Final answer
Twice the number of Dina's marbles when added to five equals seventy five. How many are Dina's marbles?	Four times the number of Bobi's marbles when added to two equals fifty. How many are Bobi's marbles?	12
Answer: Step 1: Translate the sentence into an equation • Identify the keywords. These are underlined in the sentence above. • The variable is: *the number of Dina's marbles* • Give a symbol to the variable, say it is: p • The equation is $2 \times p + 5 = 75$ or it can be written $2p + 5 = 75$ Step 2: Solve the equation $2p + 5 = 75$ $2p + 5-\mathbf{5} = 75-\mathbf{5}$ [subtract **5** from both sides] $2p = 70$ $\frac{2p}{2} = \frac{70}{2}$ [divide both sides by **2**] $p = 35$ Step 3: Make a conclusion Hence, *the number of Dina's marbles is* 35.	Answer: Step 1: Step 2: Step 3:	

Received March 31, 2016
Revision received October 29, 2016
Accepted November 1, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 5, 680–693

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000159

Developmental Change in the Influence of Domain-General Abilities and Domain-Specific Knowledge on Mathematics Achievement: An Eight-Year Longitudinal Study

David C. Geary, Alan Nicholas, Yaoran Li, and Jianguo Sun
University of Missouri

The contributions of domain-general abilities and domain-specific knowledge to subsequent mathematics achievement were longitudinally assessed (n = 167) through 8th grade. First grade intelligence and working memory and prior grade reading achievement indexed domain-general effects, and domain-specific effects were indexed by prior grade mathematics achievement and mathematical cognition measures of prior grade number knowledge, addition skills, and fraction knowledge. Use of functional data analysis enabled grade-by-grade estimation of overall domain-general and domain-specific effects on subsequent mathematics achievement, the relative importance of individual domain-general and domain-specific variables on this achievement, and linear and nonlinear across-grade estimates of these effects. The overall importance of domain-general abilities for subsequent achievement was stable across grades, with working memory emerging as the most important domain-general ability in later grades. The importance of prior mathematical competencies on subsequent mathematics achievement increased across grades, with number knowledge and arithmetic skills critical in all grades and fraction knowledge in later grades. Overall, domain-general abilities were more important than domain-specific knowledge for mathematics learning in early grades but general abilities and domain-specific knowledge were equally important in later grades.

Educational Impact and Implications Statement
The current study identifies the factors that influence students' grade-to-grade gains in mathematical competencies from school entry through middle school. The key educational finding is that the importance of students' prior mathematical knowledge for further gains in this knowledge increases across grades. The implication is that interventions focused on improving domain-specific skills, such as fractions knowledge, will likely yield longer-term gains in mathematics achievement than will interventions focused on domain-general abilities, such as working memory.

Keywords: domain-general abilities, domain-specific knowledge, mathematics achievement, mixed functional data analysis, longitudinal study

There is consensus that a combination of domain-general abilities, such as intelligence and working memory, and domain-specific knowledge contribute to academic and occupational learning, but their relative contribution to this learning is debated, including whether these contributions change over time or level of expertise (e.g., Ferrer & McArdle, 2004; Gustafsson & Undheim, 1992; Schmidt & Crano, 1974; Von Aster & Shalev, 2007). These are in fact long-standing issues in developmental, differential, and educational psychology (Ackerman, 2000; Ferrer & McArdle, 2004; Fuchs, Geary, Fuchs, Compton, & Hamlett, 2016; Geary, 2011; Gustafsson & Undheim, 1992; Schmidt & Crano, 1974; Von Aster & Shalev, 2007). For instance, Cattell's (1987) influential investment theory focused on the importance of fluid intelligence (abstract problem solving), in combination with interests and personality, in the development of crystallized intelligence, including

David C. Geary, Department of Psychological Sciences, and Interdisciplinary Neuroscience Program, University of Missouri; Alan Nicholas, Department of Statistics, University of Missouri; Yaoran Li, Department of Educational, School, and Counseling Psychology, University of Missouri; Jianguo Sun, Department of Statistics, University of Missouri.

David C. Geary acknowledges support from Grants R01 HD38283 and R37 HD045914 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development and DRL-1250359 from the National Science Foundation. We thank Mary Hoard, Lara Nugent, Linda Coutts, Drew Bailey, Sarah Becktell, Erica Bizub, Stephen Cobb, Morgan Davis, Lindsay Dial, James Dent, Margery Gurwit, Jennifer Hoffman, Stacey Jones, Jared Kester, Kristine Kuntz, Kelly Regan, Nicole Reimer, Laura Roider, Hiba Syed, Kyle Stiers, Jonathan Thacker, Leah Thomas, Erin Twellman, Tessa Vellek, Alex Wilkerson, Erin Willoughby, and Melissa Willoughby for help on various aspects of the project.

Correspondence concerning this article should be addressed to David C. Geary, Department of Psychological Sciences, University of Missouri, Columbia, MO 65211-2500. E-mail: gearyd@missouri.edu

domain-specific knowledge (see also Ackerman & Beier, 2006); ". . . this year's crystallized ability level is a function of last year's fluid ability level—and last year's interest in school work" (Cattell, 1987, p. 139). Other differential and educational psychologists have noted that level of domain-specific knowledge can influence further gains in this knowledge (Ackerman, 2000; Sweller, 2012; Tricot & Sweller, 2013; Thorsen, Gustafsson, & Cliffordson, 2014).

The results from associated empirical studies have been mixed, however. Schmidt and Crano (1974) found that intelligence measured in 4th grade predicted gains on a composite measure of academic achievement through 6th grade, controlling prior achievement, consistent with Cattell's (1987) argument. However, the pattern was found only for students from middle socioeconomic status (SES) households, not for students from lower SES households. Ferrer and McArdle (2004) found that fluid intelligence predicted gains in mathematics achievement throughout childhood and adolescence but did not assess if the magnitude of this relation varied across grades. Moreover, once intelligence, general knowledge, and performance in nonmathematical academic areas were controlled, prior mathematics achievement was inversely related to subsequent achievement. Gustafsson and Undheim (1992), in contrast, found a strong positive relation between a composite of academic skills in 6th grade and academic skills in 9th grade, but intelligence in 6th grade did not predict gains in academic skills through 9th grade, above and beyond the influence of prior skills.

The mixed results are related in part to use of different domain-general and domain-specific measures across studies, as well as different age ranges and different analytic techniques. Assessments of these relations were initially based on cross-lagged correlations (e.g., Schmidt & Crano, 1974), whereby the relation between domain-general abilities and domain-specific knowledge at one age predicts abilities and knowledge at a later age, but critiques of these methods (Kenny, 1975; Rogosa, 1980) resulted in the development and use of more sophisticated autoregressive cross-lagged models (Hamaker, Kuiper, & Grasman, 2015; Preacher, 2015). These models provide greater flexibility in the estimation of time-related change in the influence of domain-general abilities and domain-specific knowledge on the outcome of interest. In one recent study in which these methods were used, Lee and Bull (2016) found a stable across-grade influence of working memory on subsequent-grade mathematics achievement and that the importance of prior mathematics achievement increased across grades.

Functional data analysis (FDA) may provide an even more flexible approach to the study of longitudinal relations because it is not constrained by the parametric assumptions of cross-lagged and related models; indeed, the latter may be considered subsets of FDA (Müller, 2009). The mixed FDA approach has been successfully applied to many areas including image analysis and signal analysis as well as estimation of growth curves. Particularly, for the estimation of growth curves, it allows one to capture the underlying shape of the true curve much more accurately than other methods (Ramsay, Hooker, & Graves, 2009; Wang, Chiou, & Müller, 2016). In the context of this study, the FDA approach enabled the simultaneous assessment of domain-general and domain-specific effects on gains in children's mathematics achievement from 2nd to 8th grade, inclusive, and linear and nonlinear changes in the relative contributions of these effects across grades.

As with other academic domains, the relative contributions of domain-general abilities and domain-specific knowledge to subsequent mathematics achievement are not fully understood and may vary across grade, level of student knowledge, and mathematical content (Bailey, Watts, Littlefield, & Geary, 2014; Friso-van den Bos, van der Ven, Kroesbergen, & van Luit, 2013; Fuchs et al., 2016; Geary, 2011; Lee & Bull, 2016; Von Aster & Shalev, 2007; Watts et al., 2015). Identifying the grade-to-grade contributions of domain-general and domain-specific effects and changes in the relative magnitude of these effects will contribute significantly to our understanding of the factors that drive children's mathematical development and will provide insights into when and where to target interventions to improve this development.

Domain-General Abilities, Domain-Specific Knowledge, and Mathematics Achievement

Intelligence and one or more components of working memory are the most frequently studied domain-general predictors of mathematics achievement (e.g., Fuchs et al., 2016; Geary, 2011; Lee & Bull, 2016; Van de Weijer-Bergsma, Kroesbergen, & Van Luit, 2015), although reading- and language-related competencies are included in some studies (LeFevre et al., 2010; Watts, Duncan, Siegler, & Davis-Kean, 2014), as are noncognitive constructs, such as mathematics self-concept (Watts et al., 2015). Among these measures, the most consistent effects are found for intelligence and the updating component of working memory (i.e., holding information in mind while processing other information; Deary, Strand, Smith, & Fernandes, 2007; Friso-van den Bos et al., 2013; Geary, 2011; Lee & Bull, 2016; Östergren & Träff, 2013; Siegler et al., 2012), although the strength of these effects can vary across grades and with the novelty and complexity of the mathematics domain being assessed (e.g., Fuchs et al., 2010; Lee & Bull, 2016). On the basis of these findings, we included 1st grade measures of intelligence and the updating component of working memory (or central executive) as domain-general abilities. Fluid intelligence and competence at updating improve over the timeframe assessed here (e.g., Fry & Hale, 1996; Gathercole, Pickering, Ambridge, & Wearing, 2004; Li & Geary, 2013), but individual differences in these domains are at least moderately stable (Mazzocco & Kover, 2007; Sameroff, Seifer, Baldwin, & Baldwin, 1993; Thorndike, 1933).

We also included prior grade word reading achievement as a domain-general effect. Word reading is obviously a domain-specific skill, but may also tap domain general abilities. In the traditional psychometric literature, all cognitive and academic tests share common variance, typically attributed to intelligence, as well as unique test-specific variance (Carroll, 1993). Fuchs et al. (2016), as an example, found that children's early fluency with identifying letters and high-frequency words predicted subsequent word reading and arithmetic achievement, but the magnitude of the effect was 3.5-fold higher for reading than arithmetic (see also Chu, vanMarle, & Geary, 2016). The effect for arithmetic may reflect the operation of domain-general mechanisms, whereas that for reading reflects these mechanisms and the influence of domain-specific skills. There is some evidence that this domain-general

ability is not intelligence or working memory but rather the ease with which students form arbitrary visual-verbal associations (Koponen et al., 2013). This would explain the often-found relation between basic reading skills and memorization of arithmetic facts (e.g., Hecht, Torgesen, Wagner, & Rashotte, 2001), both of which may reflect the functional integrity of the domain-general hippocampal-dependent memory system (Qin et al., 2014).

Studies of the influence of earlier domain-specific knowledge on the acquisition of later knowledge are either based on the relation between mathematics achievement scores in one grade and scores on the same or a similar measure in a later grade (Bailey et al., 2014; Duncan et al., 2007; Lee & Bull, 2016), or the relation between specific mathematical competencies in one grade and achievement in a later grade. The key mathematical knowledge identified in the latter studies includes an understanding of numbers and basic arithmetic in the early grades (Clements, Sarama, Wolfe, & Spitler, 2013; Jordan, Kaplan, Ramineni, & Locuniak, 2009; Koponen, Salmi, Ekland, & Aro, 2013), and fractions in later grades (Bailey, Hoard, Nugent, & Geary, 2012; Siegler et al., 2012). Accordingly, in separate analyses domain-specific knowledge was indexed by prior grade mathematics achievement or by several measures of prior grade number, arithmetic, or fractions knowledge that have been shown to predict later mathematics achievement (Bailey et al., 2012; Booth & Siegler, 2006; Geary, 2011; Geary et al., 2013). For the latter, we assessed children's number knowledge and their basic arithmetic skills, beginning in 1st grade, and used these to predict mathematics achievement in the subsequent year; specifically, 2nd to 8th grade, inclusive. We also used a 6th grade measure of fractions knowledge to predict concurrent mathematics achievement (we did not have a 5th grade fractions assessment) and achievement in 6th, 7th, and 8th grades. The inclusion of three grades allowed us to assess nonlinear effects for this measure.

Current Study

As noted, we applied longitudinal mixed FDA techniques (Guo, 2002; Liu & Guo, 2011) to data from a unique longitudinal study to simultaneously estimate the contributions of domain-general abilities and domain-specific knowledge on mathematics achievement from 2nd to 8th grade, inclusive. The models also allowed us to estimate across-grade linear and nonlinear changes in these effects, across-grade changes in the contribution of individual domain-general and domain-specific variables, and across-grade change in the overall contribution of domain-general abilities and domain-specific knowledge on subsequent mathematics achievement. In all, the statistical approach and unique longitudinal data set that included assessment of key domain-general abilities and key domain-specific knowledge provided a more nuanced picture of stability and change in the influence of domain-general and domain-specific effects on children's mathematics achievement than previous studies and thus informs approaches to improve this achievement.

Method

Participants. All kindergarten children from 12 elementary schools that serve families from a wide range of socioeconomic backgrounds were invited to participate. Parental consent and child assent were received for 37% (n = 311) of these children, and 288 of them completed the first year of testing (see Geary, Hoard, Nugent, & Bailey, 2012). The 167 children (78 boys) used in the current analyses completed all annual achievement assessments and the vast majority of mathematical cognition tasks through the end of 8th grade. Table 1 shows the measures used in the analyses or reported elsewhere in the text and the grades in which they were administered.

The intelligence of the sample used in the analyses was average (M = 102, SD = 15), as was their mathematics achievement at the end of kindergarten (M = 103, SD = 13) and 8th grade (M = 94, SD = 18). Their reading achievement was high average at the end of kindergarten (M = 113, SD = 14) and average at the end of 8th grade (M = 101, SD = 12). The children who did not complete all of these measures (n = 121) had lower intelligence (M = 95, SD = 14), $F(1, 286) = 16.04$, $p < .0001$, and lower mathematics achievement at the end of kindergarten (M = 99, SD = 14), $F(1, 286) = 6.24$, $p = .013$, than the sample used here; however, their kindergarten reading scores (M = 110, SD = 15) did not differ, $F(1, 286) = 2.96$, $p = .0865$. Five percent of the sample identified as Hispanic. Seventy-seven percent identified as White, 4% as

Table 1
Measures Used in Analyses or Reported in Text and Grade of Administration

Variable	Grade								
	K	1	2	3	4	5	6	7	8
Mathematics achievement	x	x	x	x	x	x	x	x	x
Domain-general predictors									
Intelligence		x							
Central executive		x				x			
Reading achievement	x	x	x	x	x	x	x	x	x
Domain-specific predictors									
Number sets test		x	x	x	x	x	x	x	
Number line		x	x	x	x	x	x	x	
Addition retrieval		x	x	x	x	x	x	x	
Addition decomposition		x	x	x	x	x	x	x	
Fractions comparison test							x		

Black, 6% as Asian, 8% as mixed race, and 5% as other or unknown. The sample averaged 6 years 2 months of age (SD = 4 months) at the kindergarten achievement assessment and 14 years 2 months (SD = 4 months) at the 8th-grade achievement assessment.

Participants' parents were asked to complete a survey that included items on their education level, income, and government assistance. Complete or partial information was available for the families of 149 participants. Of these parents, 4% had some schooling but no General Education Development certificate (GED) or high school degree; 24% had a high school diploma or GED; 4% had some college, technical school, or an associate's degree; 39% had a bachelor's degree; and 29% had a postgraduate degree. The total household income was: $0–$25k (6%), $25k–$50k (22%), $50k–$75k (19%), $75k–$100k (14%), $100k–$150k (22%), $150k or more (16%). Six percent of parents reported receiving food stamps; 1% reported receiving housing assistance. We used the education level of the students' primary caregiver as an indictor of family SES; specifically, 4-year college degree (n = 101), high school diploma (n = 48), and no information (n = 18). The intelligence of children from each of these family types was in the average range. The IQ of children from high school households (M = 95, SD = 15) was lower than that of children from college (M = 105, SD = 14) and no-information (M = 103, SD = 17) households ($ps < .05$); the two latter groups did not differ ($p > .50$). The kindergarten mathematics achievement of children from college households (M = 106, SD = 13) was higher than that of children from high school (M = 98, SD = 9) and no-information (M = 98, SD = 12) households ($ps < .005$); the two latter groups did not differ ($p > .50$). At the end of 8th grade, the mathematics achievement of children from high school households (M = 86, SD = 17) was lower than that of children from college households (M = 98, SD = 18, $p < .0002$); the children from no-information households (M = 92, SD = 14) did not differ from either group ($ps > .15$).

In preliminary analyses, the parental education and income categories were included as covariates (n = 149) but did not substantively change our results with inclusion of the domain-general variables. Thus, these were not included in the final analyses, which allowed us to use data from all 167 children. We include the information here to provide a more thorough description of our sample.

Mathematics achievement. Mathematics achievement was assessed with the Numerical Operations subtest of the Wechsler Individual Achievement Test–II: Abbreviated (WIAT-II; Wechsler, 2001). The easier items include number discrimination, rote counting, number production, and basic arithmetic operations. More difficult items include rational numbers and simple algebra and geometry problems solved with pencil-and-paper. Spearman-Brown reliability estimates for the age ranges assessed here range from .87 to .96 (Mdn = .91; Wechsler, 2001).

Domain-general measures.

Intelligence. Full-scale IQ was estimated using the Vocabulary and Matrix Reasoning subtests of the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999).

Central executive. The central executive was assessed using three dual-task updating subtests of the Working Memory Test Battery for Children (WMTB-C; Pickering & Gathercole, 2001). Listening Recall requires the child to determine if a sentence is true or false and then to recall the last word in a series of sentences. Counting Recall requires the child to count a set of 4, 5, 6, or 7 dots on a card and then to recall, in order, the number of dots counted on each card at the end of that series of cards. Backward Digit Recall is a standard format backward digit span task. The subtests consist of span levels ranging from one to six or one to nine items to remember, and each span level has six trials. Failing three trials at one span level terminates the subtest, and passing four trials moves the child to the next level. The total number of trials answered correctly was used as the central executive measure, because these scores are more reliable than span scores (α = .77). To keep the assessment time consistent with our IQ measure, we used 1st grade scores as the domain-general predictor. The students were administered the same measure in 5th grade and the across grade correlation (r = .61, $p < .0001$), indicated at least moderate stability in individual differences in the central executive.

Reading achievement. Reading achievement was assessed with the Word Reading subtest of the WIAT-II (Wechsler, 2001). The easier items include matching and identifying letters, rhyming, beginning and ending sounds, and more difficult items assess accuracy of reading increasingly difficult words. Spearman-Brown reliability estimates for the age ranges assessed here range from .95 to .98 (Mdn = .97) (Wechsler, 2001).

Domain-specific mathematical cognition measures. As noted, we included specific measures of number skills, arithmetical competencies, and knowledge of fractions concepts that have been shown to predict concurrent or later mathematics achievement in previous studies (e.g., Bailey et al., 2012; Booth & Siegler, 2006; Clements et al., 2013; Jordan et al., 2009; Koponen et al., 2013; Siegler et al., 2012). These included fluency in accessing and manipulating the quantities associated with Arabic numerals (number sets test), accuracy in placing numerals on the number line (number line estimation), knowledge of basic addition facts and ability to decompose numbers into smaller sets to solve more complex addition problems (addition strategy choices), and fluency in determining the larger of two fractions (fractions comparison test).

Number sets test. Two types of stimuli are used: objects (e.g., squares) in a 1/2-in. square and an Arabic numeral (18-pt. font) in a 1/2-in. square. Stimuli are joined in domino-like rectangles with different combinations of objects and numerals. These dominos are presented in lines of five across a page. The last two lines of the page show three 3-square dominos. Target sums (5 or 9) are shown in large font at the top the page. On each page, 18 items match the target; 12 are larger than the target; six are smaller than the target; and six contain 0 or an empty square.

The tester begins by explaining two items matching a target sum of 4; then, uses the target sum of 3 for practice. The measure is then administered. The child is told to move across each line of the page from left to right without skipping any; to "circle any groups that can be put together to make the top number, 5 (9)"; and to "work as fast as you can without making many mistakes." The child has 60 s per page for the target 5; 90 s per page for the target 9. Time limits were chosen to avoid ceiling effects and to assess fluent processing of quantities.

The variable used here was based on the d' measure; specifically, [Z Hits–Z False Alarms] × [Maximum Reaction Time (RT)/Actual RT]. Thus, the scores of children who completed the test in less than

the maximum time were adjusted upward. The adjustment was made because most of the children completed the test in less than the maximum time in later grades. The adjustment enabled us to maintain the sensitivity of the test, despite faster processing times across grades. The d' score appears to capture the speed and accuracy with which children can access the magnitudes associated with whole numbers and their implicit or explicit understanding of arithmetic (Moore, vanMarle, & Geary, 2016) and is highly reliable in all grades; the Spearman-Brown reliability estimates ranged from .89 to .92 (*Mdn* = .90), based on the correlation between performance on the 5 and 9 target sums.

Number line estimation. Across grades, two different number line tasks were administered. The 1st to 5th grade, inclusive, task was a 0 to 100 number line. Here, a series of 24, 25-cm number lines containing a blank line with the two endpoints (0 and 100) was presented, one at a time, to the child with a target number (e.g., 43) in a large font printed above the line. The child's task was to mark the line where the target number should lie (Siegler & Booth, 2004); a pencil-and-paper version was used in first grade and a computerized version, where the child used the mouse to mark the line, thereafter. The procedure was the same, but the task was a 0 to 1000 number line in 6th and 7th grade. The measure was the percent absolute error, because this correlates with later mathematics achievement (Booth & Siegler, 2006) and captures children's understanding of the line (Siegler et al., 2011) regardless of the strategies used to guide the placements (Rouder & Geary, 2014); αs = .61 to .84 (*Mdn* = .72).

Addition strategy choices. Fourteen simple addition problems and six more complex problems were horizontally presented, one at a time, at the center of a computer monitor (using flash cards in 1st grade). The simple problems consisted of the integers 2 through 9, with the constraint that the same two integers (e.g., 2+2) were never used in the same problem; 1/2 of the problems summed to 10 or less and the smaller valued addend appeared in the first position for 1/2 of the problems. The complex items were 16+7, 3+18, 9+15, 17+4, 6+19, and 14+8.

The child was asked to solve each problem (without pencil-and-paper) as quickly as possible without making too many mistakes. It was emphasized that the child could use whatever strategy was easiest to get the answer, and was instructed to speak the answer into a microphone that was interfaced with the computer that in turn recorded RT from onset of problem presentation to microphone activation. After solving each problem, the child was asked to describe how they got the answer. Using the child's description and the experimenter's observations, the trial was classified as counting fingers, verbal counting, retrieval, and decomposition. The combination of experimenter observation and child reports has proven to be a useful measure of children's strategy choices (Siegler, 1987). The validity of this information is supported by RT patterns; finger-counting trials have the longest RTs, followed respectively by verbal counting, decomposition, and direct retrieval (Geary, Hoard, & Bailey, 2012).

We used two variables from these tasks. The first was the frequency with which children correctly retrieved answers to the simple addition problems. This variable indexed their relative mastery of basic facts. The second was the frequency with which decomposition was correctly used to solve the more complex problems; for instance, to solve 7+16, first decomposing 7 into 3 and 4 and then adding 16+4 and then 20+3. Use of this strategy reflects children's conceptual understanding that numbers are composed of subsets of smaller numbers (Geary, Hoard, Byrd-Craven, & DeSoto, 2004).

Fractions comparison test. The 16 items require the child to circle the larger of two presented fractions in 120 s (Geary et al., 2013). This task consists of four types of comparisons. The first type presents two fractions with a constant numerator but different denominators (e.g., 1/5 vs. 1/9), which assesses the child's understanding of the inverse relation between the denominator and the fraction value. In the second type, both numerators and denominators differ and the fraction with the larger value always has the larger numerator and smaller denominator (e.g., 3/10 vs. 2/12). If the child focuses on the numerators and chooses the larger one, the child will always be correct. If the child focuses on the denominators, in contrast, and chooses the larger one, the child will always be incorrect. The numerators in each comparison pair have a ratio of 1.5 and denominators have ratios between 1.1 and 1.25; these ratios were chosen based on the Weber function for magnitude discrimination in adolescents (Halberda & Feigenson, 2008). In the third type of comparison, numerators and denominators are reversed (e.g., 3/2 vs. 2/3) to assess whether the child understands that the larger fraction should have the larger numerators and the smaller denominators. The comparisons in the fourth type involve a fraction with a 1/2 value as an anchor and the other fraction close to 1 (e.g., 20/40 vs. 8/9). A child who understands fraction magnitudes should be able to quickly identify the 1/2 fraction and choose the other one.

The four types of comparisons were designed to examine whether the children conceptually understand the meanings of the numerator, the denominator, and the value of a fraction as a whole. The child received 1 point for circling the correct answer in each pair. Scores were summed across item type to create a single composite score (α = .83).

Procedure.

Assessments. The achievement measures were administered in the spring semester of each academic year and the WASI was administered in the spring of first grade. The addition strategy and number sets tasks were administered in the fall of each academic year. The number line measure was administered in the fall of first grade and in the spring of subsequent academic years to accommodate time constraints in the fall assessment. The fractions comparison task was administered in the spring semester in 6th grade. The majority of children were tested in a quiet location at their school site, and occasionally on the university campus or in a mobile testing van. Testing in the van occurred for children who had moved out of the school district or to a nonparticipating school and for administration of the WMTB-C (e.g., on the weekend or after school). The mathematical cognition and achievement assessments required about 40 min and the WMTB-C about 60 min per assessment.

Analyses. For the 167 students used in these analyses, the 1.3% of missing data were imputed using the Multiple Imputation procedure in SAS (SAS, 2014), using the EM algorithm with multivariate normality assumption. All variables were standardized ($M = 0$, $SD = 1$) at each grade level to provide a common metric for estimating across-grade change in the relative contribution of domain-general abilities and domain-specific knowledge on subsequent mathematics achievement, following the mixed FDA discussed in Guo (2002) and in Liu and Guo (2011). The method allows for both the estimation of regression effects that are linear

functions of time (grade) from longitudinal data and the estimation of nonlinearity in these functions of time. For n subjects assessed across multiple grades, subject i is observed at grades, $g_{i1} < \ldots < g_{imi}$, $i = 1, \ldots, n$. The outcome of interest is mathematics achievement or y, such that y_{ij} denotes the observed value on subject i at grade g_{ij}, $j = 1, \ldots, m_i$, $i = 1, \ldots, n$. Then the FDA model assumes that y_{ij} can be described by

$$y_{ij} = X_{ij}\beta(g_{ij}) + Z_{ij}\alpha_i(g_{ij}) + e_{ij}, \quad (1)$$

whereby $\beta(g) = \{\beta_1(g), \ldots, \beta_p(g)\}$represents a set of fixed functions (i.e., domain-general and domain-specific regression effects), and $\alpha_i(g) = \{\alpha_{1i}(g), \ldots, \alpha_{qi}(g)\}$ represents a set of random functions (i.e., individual subject effects), X_{ij}, Z_{ij} represent the two matrices of the study design, and e_{ij} is the error term. Liu and Guo (2011) suggest to represent both fixed and random functions by linear splines, which are piecewise linear polynomials joined end-to-end at the joints (grades). For each individual variable, the across-grade estimates are simultaneously and jointly calculated.

One major advantage is that the method allows estimation of the curve for each individual variable and results in smooth and more natural estimates for the overall regression functions, that is, the grade-over-grade estimates of the relation between prior domain-general abilities and domain-specific knowledge on subsequent mathematics achievement. The estimates were obtained using the %fmixed SAS macro, following Liu and Guo (2011). All substantive models were variations of Equation 1. The two domain-specific models, Model 1 and Model 2 were represented by,

$$NO_i(g) = \beta_1(g)NO_i(g-1) + \alpha_{1i}(g) + e_i(g), \quad (2)$$

$$NO_i(g) = \beta_1(g)Ret_i(g-1) + \beta_2(g)Dec_i(g-1) + \beta_3(g)NS_i(g-1) + \beta_4(g)NL_i(g-1) + \alpha_{1i}(g) + e_i(g), \quad (3)$$

whereby NO is Numerical Operations scores, and $\beta_1(g)NO_i(g-1)$ represents the prediction of these scores by Numerical Operations scores from the preceding grade. Model 1 (Equation 2) thus estimates the effects of prior-grade mathematics achievement on subsequent mathematics achievement. For Model 2 (Equation 3), the effect of Numerical Operations scores are replaced by preceding grades' fact retrieval (*Ret*), use of decomposition (*Dec*), number sets fluency (*NS*), and number line accuracy (*NL*). Model 2 thus allows for the estimation of domain-specific effects for all of these variables unadjusted for domain-general effects and grade-to-grade changes in the magnitudes of each of these effects. Comparison of the fit of Model 1 and Model 2 assesses whether broad achievement measures or specific quantitative knowledge are the better indicators of domain-specific effects.

As noted, domain-general abilities included 1st grade intelligence (*IQ*) and 1st grade central executive (*CE*) scores, as well as reading scores (*Read*) from the prior grade, as represented by Model 3 (see Equation 4). Model 3 thus provides an overall assessment of the importance of these domain-general abilities on subsequent mathematics achievement, unadjusted by domain-specific effects. The fourth model (equation not shown) combined all terms from Models 2 and 3, and thus provides the critical adjusted simultaneous estimate of domain-specific and domain-general effects, and grade-to-grade changes in the magnitude of these effects, on subsequent mathematics achievement. The first through fourth models provide estimates for domain-specific and domain-general effects unadjusted or adjusted from 2nd to 8th grade, inclusive. The fifth model (equation not shown) added the fractions comparison measure at 6th grade to the model for combined effects, but only estimated effects from 6th to 8th grade, inclusive.

$$NO_i(g) = \beta_1(g)IQ_i + \beta_2(g)Read_i(g-1) + \beta_3(g)CE_i(g) + \alpha_{1i}(g) + e_i(g), \quad (4)$$

Results

Independent or unadjusted domain-specific and domain-general effects. Table 2 shows the variance components for the fixed and random effects from all of the models, as well as the mean R^2 over grades for each model. The spline variance components in particular indicate the extent to which across-grade estimates deviate from a straight line, with 0 estimates indicating linear trends and >0 estimates indicating nonlinear trends, or spline. The grade-to-grade estimates of fixed domain-specific effects for Numerical Operations (Model 1) and the four mathematical cognition measures (Model 2) are shown in Figure 1 (Panel A).

When no other variables are included in the model, the importance of prior grades' Numerical Operations scores on subsequent scores increases across grades, suggesting a gradual increase in the importance of domain-specific knowledge. It is also clear from Figure 1 that the importance of each individual mathematical cognition variable for predicting subsequent achievement is smaller than prior Numerical Operations scores, but in combination the mathematical cognition variables explain substantively more variation in subsequent achievement (mean R^2 = .822, Table 2) than does Numerical Operations (mean R^2 = .682). The latter indicates that domain-specific effects are better represented by the combination of mathematical cognition scores than by scores on prior mathematics achievement tests, and thus the mathematical cognition scores were included in combined models.

As shown in Table 2, estimation of the three domain-general effects (Model 3) explained a substantial amount of variance in subsequent Numerical Operations scores, mean R^2 = .814. The

Table 2
Variance Components Associated With Fixed Effects

	Model				
	1	2	3	4	5
Fixed-effect spline variance component					
IQ			.0000	.0000	.0000
Reading (g-1)			.0000	.0000	.0000
Central executive			.8860	.3981	.0919
Numerical operations (g-1)	.0000				
Number sets (g-1)		.0000		.0000	.0000
Retrieval(g-1)		.0000		.0000	.0000
Decomposition(g-1)		.4709		.8808	.0000
Number line (g-1)		.0000		.0000	.0000
Fractions comparison		—			.0000
Random-effect variance component					
Intercept	.2581	.3661	.3309	.2317	.3144
Slope	.0353	.2325	.5232	.2569	.0973
Spline	.0000	6.6547	10.1853	6.6063	.0000
Residual VC	.3587	.2405	.2544	.2341	.1731
Mean R^2 over grade	.682	.822	.814	.828	.784

Note. g = grade; VC = variance component.

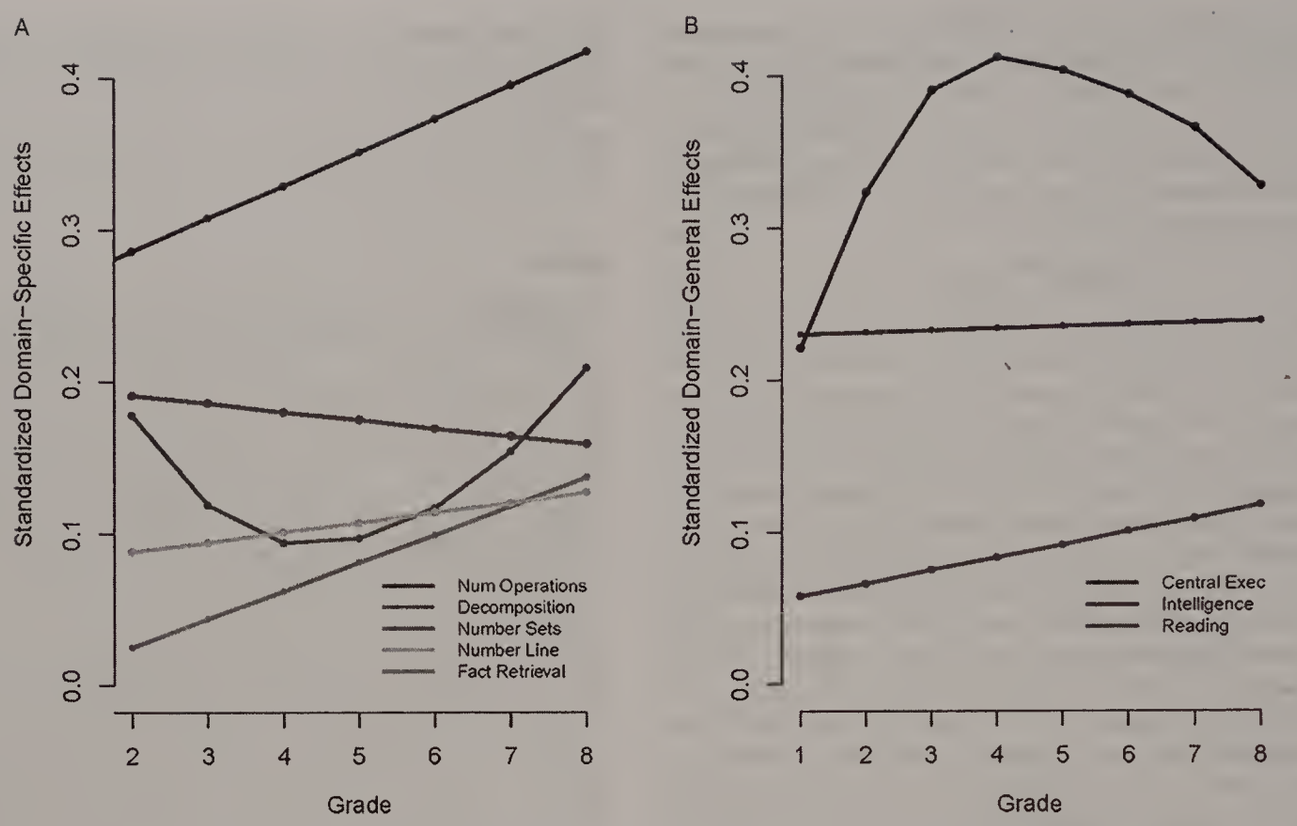


Figure 1. Standardized domain-specific effects (Panel A) from Model 1 (numerical operations) and Model 2 (mathematical cognition) and standardized domain-general effects (Panel B) from Model 3. See the online article for the color version of this figure.

effects for intelligence and reading achievement were linear across grades. The estimates for the central executive increased and then decreased across grades (Figure 1, Panel B), but the deviation from linearity was not significant ($p = .398$).

Simultaneous or adjusted domain-specific and domain-general effects. The core analyses are the simultaneous estimation of domain-specific and domain-general effects across grades in Model 4, and the determination of whether the relative magnitudes of these effects change across grades. The associated estimates are shown in Table 3 and Figure 2. None of the across-grade effects differ significantly from a straight line. As shown in Table 2 and Figure 2, there were potential nonlinear effects for the decomposition and central executive variables, but the overall deviation from a straight line was not significant for either variable ($ps > .4432$). There were nevertheless differences across the larger and smaller grade-level estimates for these two variables. The smallest effects for decomposition were for Grades 4 and 5 and the largest for Grades 2 and 8. Follow up contrasts of these grades confirmed larger effects for Grades 2 and 8 than for Grades 4 and 5 ($ps < .039$). Similar contrasts for the central executive revealed no differences between Grade 2 (smallest effect) and Grade 5 (largest effect; $p = .107$), but there was a trend for the contrast of Grades 5 and 8 ($p = .089$).

As can be seen in Table 3, the pointwise significance of the grade-level effects for the domain-specific variables is mixed through 5th grade, that is, some of the effects are significant (e.g., 1st grade number sets fluency predicting 2nd Grade Numerical Operations scores, $p < .001$), but others are not (e.g., 1st grade fact retrieval predicting 2nd Grade Numerical Operations scores, $p = .733$). After 5th grade, all of the individual domain-specific effects are significant ($ps < .038$), consistent with a gradual increase in the importance of domain-specific knowledge. With the exception of 7th Grade Word Reading scores predicting 8th Grade Numerical Operations scores ($p = .084$), all of the individual domain-general effects are significant in every grade.

To determine if there was an overall trend of increasing or decreasing effects of individual variables across 2nd to 8th grade for this combined model, we tested whether the linear slope of the across-grade fixed effect estimates was statistically different from 0; we excluded the central executive and decomposition variables because of the across-grade nonlinearity noted above. The linear slope was nonsignificant for intelligence ($p = .7248$), reading ($p = .8882$), number sets ($p = .7309$), addition fact retrieval ($p = .1102$), and number line ($p = .3783$). With the possible exception of the central executive and decomposition, the pattern suggests that despite grade-to-grade differences in the significance of individual domain-specific and domain-general variables, the magnitude of the relation between these variables and subsequent mathematics achievement was not different from constants across grades.

A similar overall pattern is evident for 6th to 8th grade, with the inclusion of the fractions comparison measure (Figure 3, Table 4). With the inclusion of the latter, the effects of intelligence ($p = .217$) and Word Reading ($p = .144$) are no longer significant by 8th grade, but the central executive remains important ($p = .007$). For the same grade, four of the five mathematical cognition measures are significant ($ps < .039$), and the individual effect for frac-

Table 3
Domain-Specific and Domain-General Effect Estimates From Model 4

Grade	Number sets (g-1)	*p*	Retrieval (g-1)	*p*	Decomposition (g-1)	*p*	Number line (g-1)	*p*
	Domain-specific effect estimates							
2	.159 (.040)	.000	.013 (.037)	.733	.179 (.051)	.000	.052 (.041)	.209
3	.155 (.032)	.000	.029 (.030)	.332	.085 (.036)	.020	.062 (.033)	.058
4	.151 (.026)	.000	.046 (.025)	.068	.047 (.034)	.175	.072 (.026)	.005
5	.148 (.024)	.000	.062 (.023)	.008	.051 (.034)	.141	.081 (.022)	.000
6	.144 (.026)	.000	.079 (.026)	.003	.073 (.035)	.038	.091 (.024)	.000
7	.140 (.032)	.000	.095 (.032)	.003	.118 (.037)	.002	.101 (.030)	.001
8	.137 (.040)	.001	.112 (.039)	.005	.192 (.051)	.000	.111 (.039)	.004
	Domain-general effect estimates							
	IQ	*p*	Central executive	*p*	Reading (g-1)	*p*		
2	.167 (.059)	.005	.208 (.066)	.002	.104 (.051)	.041		
3	.162 (.002)	.052	.266 (.055)	.000	.102 (.043)	.017		
4	.158 (.048)	.001	.299 (.053)	.000	.100 (.038)	.008		
5	.154 (.046)	.001	.305 (.052)	.000	.098 (.036)	.006		
6	.149 (.048)	.002	.293 (.052)	.000	.097 (.039)	.013		
7	.145 (.052)	.006	.266 (.055)	.000	.095 (.045)	.036		
8	.141 (.059)	.017	.219 (.065)	.001	.093 (.054)	.084		

Note. g = grade. Parenthetical values are standard errors.

tions comparison ($\beta = .206$, $p < .002$) is at least as important as that of the central executive ($\beta = .18$, $p < .007$). Examination of 6th to 8th grade change in the relation between these variables and subsequent mathematics achievement revealed a trend for a decline in the importance of intelligence ($p = .096$) and a significant increase in the importance of decomposition ($p = .015$).

The overall importance of the domain-specific and domain-general variables on subsequent mathematics achievement is

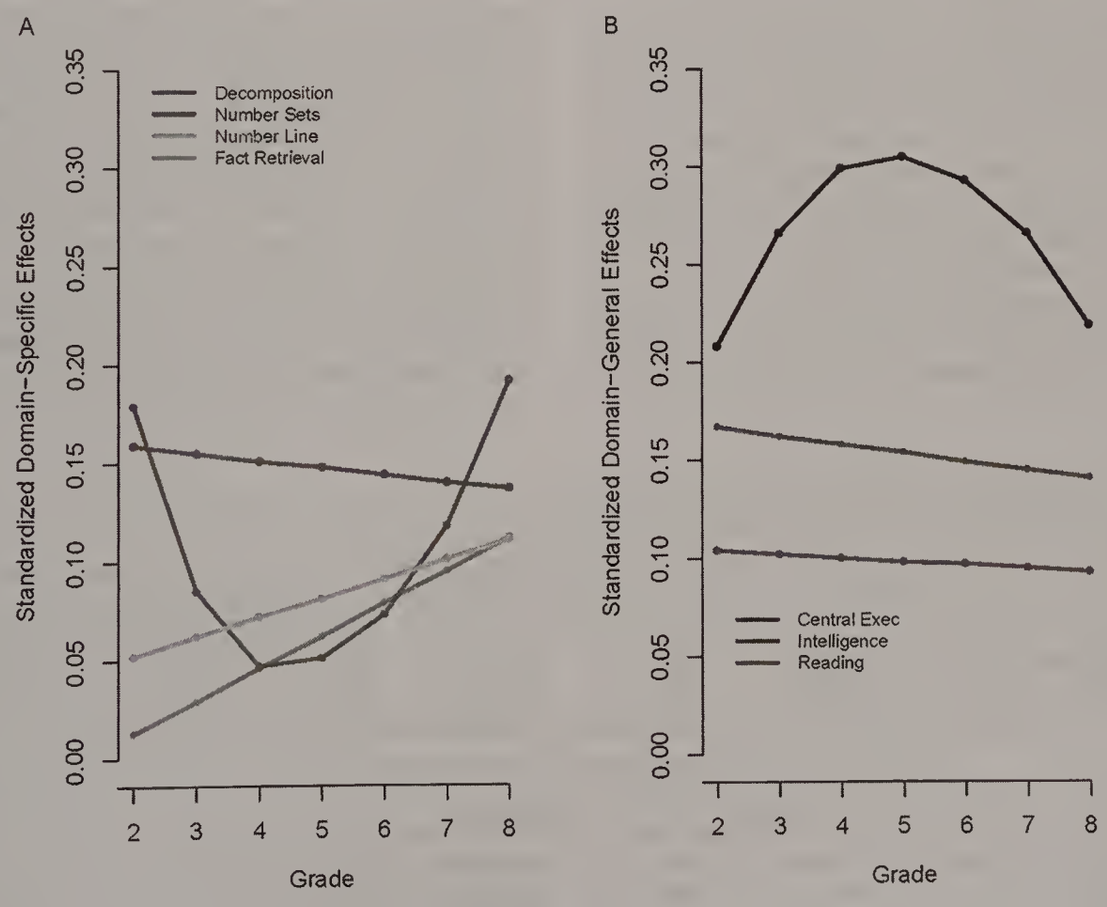


Figure 2. Standardized domain-specific (Panel A) and domain-general (Panel B) effects from 2nd to 8th grade from Model 4. See the online article for the color version of this figure.

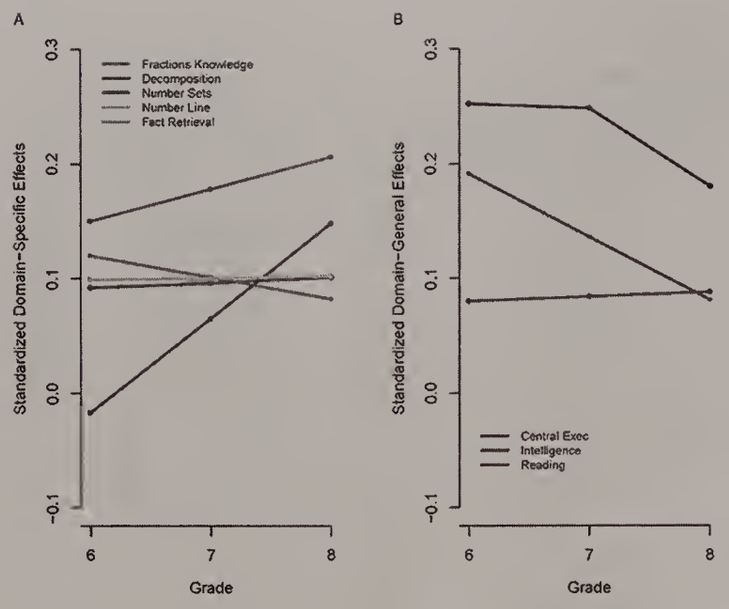


Figure 3. Standardized domain-specific (Panel A) and domain-general (Panel B) effects from 6th to 8th grade from Model 5. See the online article for the color version of this figure.

provided by the sum of the respective β estimates for each grade. As shown in Figure 4, the overall domain-general effects are relatively stable, ranging between .453 and .558. None of the adjacent grade comparisons differ significantly for these domain-general effects (*p*s > .126), nor does the contrast of the smallest (Grade 8) and largest (Grades 4 and 5) effects (*p*s > .106). The overall domain-specific effects, in contrast, are more variable, ranging from .315 to .551. Adjacent grade comparisons indicated a significant decrease in overall domain-specific effects from Grade 2 to Grade 3 ($p = .050$), but increases from Grade 6 to Grade 7 ($p = .034$), and Grade 7 to Grade 8 ($p = .024$). Moreover, the contrast of the smallest (Grade 4) to largest (Grade 8) overall domain-specific effect was highly significant ($p = .007$). There is a trend for larger overall domain-general than domain-specific effects in Grade 3 ($p = .067$) and significant differences in Grades 4 ($p = .014$) and 5 ($p = .023$). As shown in Figure 4, the overall domain-specific effect exceeds the domain-general effect by 8th grade, but none of the overall differences are significant after 5th grade (*p*s > .117).

Discussion

The combination of longitudinal mixed functional data analysis and a unique data set enabled a more nuanced assessment of a long-standing question in psychology than afforded by previous studies and analytic approaches (e.g., Ackerman, 2000; Ferrer & McArdle, 2004; Fuchs et al., 2016; Geary, 2011; Gustafsson & Undheim, 1992; Schmidt & Crano, 1974; Thorsen et al., 2014; Von Aster & Shalev, 2007); specifically, the relative contributions of prior domain-specific knowledge and domain-general abilities on subsequent achievement and estimation of grade-over-grade change in the relative contribution of knowledge and abilities on this achievement. Moreover, the outcome itself, the development of mathematical competencies, is critically important for success in a wide range of jobs in the modern economy and for navigating the many now-routine activities of daily life (Bynner, 1997; Reyna, Nelson, Han, & Dieckmann, 2009), and thus the results are practically important.

Domain-general abilities. Our finding that intelligence, the central executive, and reading achievement made significant contributions to subsequent mathematics achievement in most or all grades is consistent with many previous studies (Fuchs et al., 2016; Geary, 2011; Lee & Bull, 2016; LeFevre et al., 2010; Van de Weijer-Bergsma et al., 2015; Watts et al., 2015). There are nevertheless several aspects of our results that expand on this literature. The first is that the combination of these three measures, without inclusion of the domain-specific effects, explained substantial variance in mathematics achievement, suggesting these variables captured the bulk of broadly defined domain-general abilities. This does not necessarily mean that these are the core domain-general abilities that contribute to individual differences in mathematics achievement, but rather they provide a strong proxy for a broad set of abilities that are correlated with the measures assessed here and that also contribute to mathematics achievement (e.g., Fuchs et al., 2010; Fuchs et al., 2016).

Moreover, we did not have assessments of noncognitive traits that could, in theory, also predict academic achievement (e.g., Cattell, 1987; Eccles, Vida, & Barber, 2004; Ma, 1999). Watts et

Table 4
Domain-Specific and Domain-General Effect Estimates From Model 5

Domain-specific effect estimates										
Grade	Number sets (g-1)	*p*	Retrieval (g-1)	*p*	Decomposition (g-1)	*p*	Number line (g-1)	*p*	Fractions	*p*
6	.092 (.048)	.056	.120 (.048)	.014	−.017 (.050)	.727	.099 (.047)	.034	.150 (.065)	.021
7	.096 (.036)	.007	.101 (.035)	.004	.065 (.038)	.085	.100 (.031)	.001	.178 (.057)	.002
8	.101 (.047)	.031	.082 (.045)	.071	.148 (.050)	.003	.102 (.049)	.039	.206 (.068)	.002
Domain-general effect estimates										
Grade	IQ	*p*	Central Executive	*p*	Reading (g-1)	*p*				
6	.191 (.065)	.003	.252 (.067)	.000	.080 (.063)	.203				
7	.136 (.057)	.016	.248 (.061)	.000	.084 (.050)	.095				
8	.081 (.066)	.217	.180 (.067)	.007	.088 (.060)	.144				

Note. g = grade. Parenthetical values are standard errors.

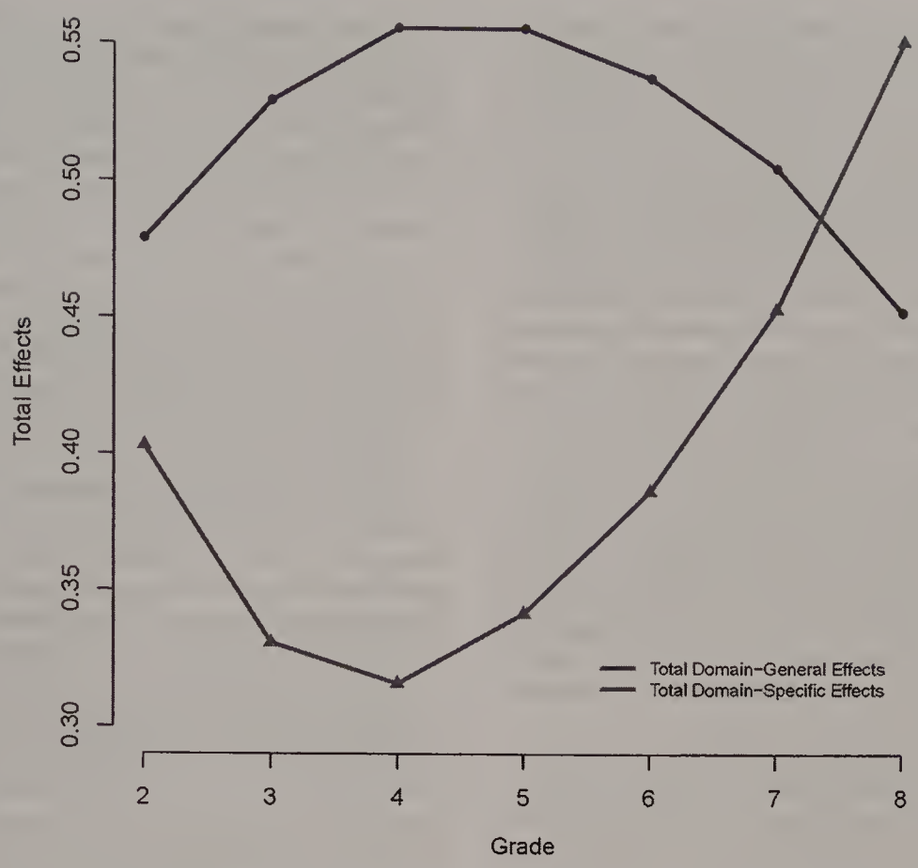


Figure 4. Total domain-specific and domain-general effects from 2nd to 8th grade. See the online article for the color version of this figure.

al. (2015), for instance, found that mathematics self-concept in 6th grade was significantly correlated with later mathematics achievement and with earlier performance on working memory measures. Moreover, mathematics self-concept predicted later mathematics achievement, controlling working memory and domain-specific division and fractions knowledge, suggesting an important noncognitive effect. As they note, however, mathematics self-concept could be a reflection of prior achievement and is related to later mathematics only because of the strong correlations between earlier and later achievement (Duncan et al., 2007). In other words, it remains to be determined if the magnitude of our specific domain-general effects would change with inclusion of noncognitive measures, such as mathematics self-concept or mathematics anxiety, that often predict later achievement.

Second, our overall results suggest stable across-grade domain-general effects on subsequent mathematics achievement. Some previous studies are also consistent with stable effects (Bailey et al., 2014), but more often than not the influence of domain-general abilities varies across grades when domain-specific variables are included as predictors (e.g., Fuchs et al., 2016; Lee & Bull, 2016; Van de Weijer-Bergsma et al., 2015). In the latter case, domain-general effects are often indirect, mediated by the relation between these abilities and prior domain-specific achievement or individual domain-specific competencies, such as arithmetic skills (e.g., Fuchs et al., 2016; Östergren & Träff, 2013). Although Cattell's (1987) ideas have not been incorporated into this literature, the associated results are consistent with his argument that measures of crystallized abilities or domain-specific knowledge reflect, in part, prior levels of fluid intelligence. Our results show this same pattern, whereby domain-general abilities influence subsequent mathematics achievement, and individual differences in this achievement or in specific mathematical competencies facilitate further mathematical gains.

Our domain-general measures, however, do not provide pure assessments of fluid intelligence and thus are not a direct test of Cattell's (1987) investment hypothesis. Nevertheless, the combination of standard IQ scores and working memory used here will be highly correlated with fluid abilities (e.g., Conway, Cowan, Bunting, Therriault, & Minkoff, 2002; Deary, 2000; Engle, Kane, & Tuholski, 1999; Geary, 2005), and therefore our results, though not definitive, are consistent with Cattell's hypothesis. With the inclusion of prior reading achievement in the set of domain-general variables, our results could also be interpreted as being consistent with Carroll's (1993) general intelligence that subsumes fluid intelligence and other processes that influence learning; including ease of associative learning that may underlie the relation between word reading and mathematics achievement (Chu et al., 2016; Koponen et al., 2013). Whichever way the results are framed, the key finding is that domain-general abilities, including those assessed 7 years earlier, influence mathematics achievement throughout much of formal schooling, even with control of domain-specific competencies.

Of these abilities, the central executive component of working memory (updating) emerged as particularly important. Our results for this measure are consistent with a recent cross-sequential preschool to 9th grade study in which Numerical Operations scores were predicted by prior achievement, an updating measure of working memory, and intelligence (Lee & Bull, 2016). The rela-

tion between working memory and subsequent mathematics achievement was stronger in earlier than later grades, but statistically constraining the relation to be identical across grades fitted the data nearly as well as allowing these relations to vary: The most parsimonious explanation is that the relation between working memory and subsequent achievement is stable across grades.

At the same time, Lee and Bull (2016) found that intelligence did not predict achievement in early grades, once working memory and prior achievement were controlled, but was important in 6th to 9th grade. Our results suggest a more consistent relation between intelligence and achievement across grades, but perhaps a declining influence in later grades, once fractions knowledge is included as a domain-specific effect. The differences might be related to different analytic approaches, use of different intelligence tests (they used a block design test), estimation of domain-specific effects using prior achievement versus our mathematical cognition variables, or some combination. Either way, both studies are consistent with intelligence as an important domain-general ability (see also Deary et al., 2007), although perhaps less important than working memory in some grades and more important in others, depending on mathematical content and students' prior knowledge.

The inclusion of prior reading achievement is not as straightforward as a domain-general ability as working memory and intelligence, although use of reading- and language-related measures in similar studies is common (e.g., Fuchs et al., 2016; LeFevre et al., 2010; Watts et al., 2015). Still, word reading should have little if any direct effect on solving problems on the Numerical Operations test, but as noted may index the ease of associative learning (Koponen et al., 2013) and functional integrity of the hippocampal-dependent memory system (Qin et al., 2014). The latter would be consistent with Cattell's (1987) 'rote learning' contributions to the development of domain-specific knowledge and is a component of Carroll's (1993) model of general intelligence. Nevertheless, further work is needed on the basic cognitive and neural mechanisms, above and beyond intelligence and working memory, that contribute to ease of learning some aspects of reading competencies and some aspects of mathematical competencies (Geary, 1993).

Domain-specific mathematical knowledge. As we found here, Lee and Bull (2016) reported that the relative importance of prior mathematics achievement on subsequent achievement increased across grades. The across-grade increase in the importance of mathematic knowledge for the learning of new mathematical knowledge, with relatively stable domain-general effects, supports instructional approaches that focus on learning domain-specific knowledge rather than teaching domain-general problem-solving competencies (Tricot & Sweller, 2013), and is consistent with individual differences in adult levels of expertise in mathematics and other domains (Ackerman, 2000; Ackerman & Beier, 2006). Our results and those of Lee and Bull (2016) also suggest that individual differences in domain-specific mathematical competencies may be a critical factor driving the greater variability among students in mathematics achievement in later than earlier grades. The implication is that addressing prerequisite domain-specific skill deficits has the potential to substantially reduce individual differences in mathematical competencies at school completion. Our results suggest these prerequisite skills include number knowledge and basic arithmetic in the early grades and fractions knowledge in later grades, consistent with other studies (e.g., Siegler et al., 2012). However, the same caveat for our domain-general measures applies here; although our measures were carefully chosen based on prior studies (Bailey et al., 2012; Booth & Siegler, 2006; Clements et al., 2013; Jordan et al., 2009; Koponen et al., 2013; Siegler et al., 2012), we cannot conclude that these measures capture all of the key domain-specific knowledge needed to progress in mathematics. It is likely that our measures are important, but other measures are likely to be just as important, and indeed the set of critical prior skills may vary to some extent across grades and with the mathematical outcome of interest. For instance, the Numerical Operations test does not include many geometry items and thus the importance of prior geometrical knowledge on future mathematics achievement could not be assessed.

Finally, a few words for our decomposition variable: Although the overall U-shaped relation between use of decomposition and subsequent-grade mathematics achievement was not significant, there was evidence that students' who frequently used decomposition had higher achievement than their peers in earlier and later grades. In early grades, only children with the most sophisticated understanding of numbers and the relations among them use decomposition with any frequency and thus its importance early in elementary school makes sense (Geary et al., 2012). The reason for the reemergence of decomposition for predicting achievement in later grades is less clear, however. Given that mean use of decomposition did not change after 4th grade (not reported here), the effect for later grades is likely due to changes in the content of the items on the Numerical Operations test; specifically, complex whole number arithmetic problems where decomposition might be a useful problem solving strategy.

Summary and Limitations

As noted, we did not include all potential domain-general and domain-specific measures in our study and thus it is possible that alternative variables might emerge in future studies. Although the data set is unique in many ways, the sample size is relatively small which may have reduced the statistical power of some of our analyses, and it is unclear how sample recruitment, attrition, and diversity (77% White) influenced our findings. The measurement of intelligence and the central executive in 1st grade and the domain-specific competencies in the grade prior to the mathematical outcome may have biased the results in favor of domain-specific competencies. We do not believe this is a strong bias, however, because Lee and Bull (2016) assessed working memory in each grade and, as noted, found the same across-grade pattern as emerged here. Moreover, if assessment timing was critical, the importance of domain-general abilities would have declined across grades, not remained stable. Of course, the data itself is correlational and does not support strong causal inferences. Follow up studies will be needed to fully assess the validity of our conclusions; such as, improvements in domain-specific skills will reduce individual differences in subsequent mathematics achievement.

Despite these limitations, the study yielded three key findings. First, the overall magnitude of domain-general effects on mathematics achievement remained constant across grades, whereas the overall magnitude of domain-specific effects increased across grades. Second and in keeping with previous studies, domain-general effects were larger than domain-specific effects in the early grades; however, the overall contributions of domain-general abilities and domain-specific

knowledge did not differ in later grades. Third, the combination of several specific measures of mathematical knowledge provided a substantively larger estimate of domain-specific effects than the more commonly used prior mathematics achievement. The later suggests that use of prior achievement scores may underestimate the importance of domain-specific knowledge for further learning in the domain, mathematics in this case.

References

Ackerman, P. L. (2000). Domain-specific knowledge as the "dark matter" of adult intelligence: Gf/Gc, personality and interest correlates. *The Journals of Gerontology Series B, Psychological Sciences and Social Sciences, 55,* 69–84. http://dx.doi.org/10.1093/geronb/55.2.P69

Ackerman, P. L., & Beier, M. E. (2006). Determinants of domain knowledge and independent study learning in an adult sample. *Journal of Educational Psychology, 98,* 366–381. http://dx.doi.org/10.1037/0022-0663.98.2.366

Bailey, D. H., Hoard, M. K., Nugent, L., & Geary, D. C. (2012). Competence with fractions predicts gains in mathematics achievement. *Journal of Experimental Child Psychology, 113,* 447–455. http://dx.doi.org/10.1016/j.jecp.2012.06.004

Bailey, D. H., Watts, T. W., Littlefield, A. K., & Geary, D. C. (2014). State and trait effects on individual differences in children's mathematical development. *Psychological Science, 25,* 2017–2026. http://dx.doi.org/10.1177/0956797614547539

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology, 42,* 189–201. http://dx.doi.org/10.1037/0012-1649.41.6.189

Bull, R., & Lee, K. (2014). Executive functioning and mathematics achievement. *Child Development Perspectives, 8,* 36–41. http://dx.doi.org/10.1111/cdep.12059

Bynner, J. (1997). Basic skills in adolescents' occupational preparation. *Career Development Quarterly, 45,* 305–321. http://dx.doi.org/10.1002/j.2161-0045.1997.tb00536.x

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511571312

Cattell, R. B. (1987). *Intelligence: Its structure, growth, and action.* Amsterdam, the Netherlands: North-Holland.

Chu, F. W., vanMarle, K., & Geary, D. C. (2016). Predicting children's reading and mathematics achievement from early quantitative knowledge and domain-general cognitive abilities. *Frontiers in Psychology, 7,* 775. http://dx.doi.org/10.3389/fpsyg.2016.00775

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *American Educational Research Journal, 50,* 812–850. http://dx.doi.org/10.3102/0002831212469270

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30,* 163–183. http://dx.doi.org/10.1016/S0160-2896(01)00096-4

Deary, I. J. (2000). *Looking down on human intelligence: From psychophysics to the brain.* New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780198524175.001.0001

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35,* 13–21. http://dx.doi.org/10.1016/j.intell.2006.02.001

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43,* 1428–1446. http://dx.doi.org/10.1037/0012-1649.43.6.1428

Eccles, J. S., Vida, M. N., & Barber, B. (2004). The relation of early adolescents' college plans and both academic ability and task-value beliefs to subsequent college enrollment. *The Journal of Early Adolescence, 24,* 63–77. http://dx.doi.org/10.1177/0272431603260919

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 102–134). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139174909.007

Ferrer, E., & McArdle, J. J. (2004). An experimental analysis of dynamic hypotheses about cognitive abilities and achievement from childhood to early adulthood. *Developmental Psychology, 40,* 935–952. http://dx.doi.org/10.1037/0012-1649.40.6.935

Friso-van den Bos, I., van der Ven, S. H., Kroesbergen, E. H., & van Luit, J. E. (2013). Working memory and mathematics in primary school children: A meta-analysis. *Educational Research Review, 10,* 29–44. http://dx.doi.org/10.1016/j.edurev.2013.05.003

Fry, A. F., & Hale, S. (1996). Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological Science, 7,* 237–241. http://dx.doi.org/10.1111/j.1467-9280.1996.tb00366.x

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Hamlett, C. L., Seethaler, P. M., . . . Schatschneider, C. (2010). Do different types of school mathematics development depend on different constellations of numerical versus general cognitive abilities? *Developmental Psychology, 46,* 1731–1746. http://dx.doi.org/10.1037/a0020662

Fuchs, L. S., Geary, D. C., Fuchs, D., Compton, D. L., & Hamlett, C. L. (2016). Pathways to third-grade calculation versus word-reading competence: Are they more alike or different? *Child Development, 87,* 558–567. http://dx.doi.org/10.1111/cdev.12474

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology, 40,* 177–190. http://dx.doi.org/10.1037/0012-1649.40.2.177

Geary, D. C. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin, 114,* 345–362. http://dx.doi.org/10.1037/0033-2909.114.2.345

Geary, D. C. (2005). *The origin of mind: Evolution of brain, cognition, and general intelligence.* Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/10871-000

Geary, D. C. (2011). Cognitive predictors of individual differences in achievement growth in mathematics: A five-year longitudinal study. *Developmental Psychology, 47,* 1539–1552. http://dx.doi.org/10.1037/a0025510

Geary, D. C., Hoard, M. K., & Bailey, D. H. (2012). Fact retrieval deficits in low achieving children and children with mathematical learning disability. *Journal of Learning Disabilities, 45,* 291–307. http://dx.doi.org/10.1177/0022219410392046

Geary, D. C., Hoard, M. K., Byrd-Craven, J., & DeSoto, M. C. (2004). Strategy choices in simple and complex addition: Contributions of working memory and counting knowledge for children with mathematical disability. *Journal of Experimental Child Psychology, 88,* 121–151. http://dx.doi.org/10.1016/j.jecp.2004.03.002

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2012). Mathematical cognition deficits in children with learning disabilities and persistent low achievement: A five-year prospective study. *Journal of Educational Psychology, 104,* 206–223. http://dx.doi.org/10.1037/a0025398

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2013). Adolescents' functional numeracy is predicted by their school entry number system knowledge. *PLoS One, 8*(1), e54651. http://dx.doi.org/10.1371/journal.pone.0054651

Guo, W. (2002). Functional mixed effects models. *Biometrics, 58,* 121–128. http://dx.doi.org/10.1111/j.0006-341X.2002.00121.x

Gustafsson, J. E., & Undheim, J. O. (1992). Stability and change in broad and narrow factors of intelligence from ages 12 to 15 years. *Journal of Educational Psychology, 84,* 141–149. http://dx.doi.org/10.1037/0022-0663.84.2.141

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the "Number Sense": The Approximate Number System in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology, 44,* 1457–1465. http://dx.doi.org/10.1037/a0012682

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods, 20,* 102–116. http://dx.doi.org/10.1037/a0038889

Hecht, S. A., Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2001). The relations between phonological processing abilities and emerging individual differences in mathematical computation skills: A longitudinal study from second to fifth grades. *Journal of Experimental Child Psychology, 79,* 192–227. http://dx.doi.org/10.1006/jecp.2000.2586

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology, 45,* 850–867. http://dx.doi.org/10.1037/a0014939

Kenny, D. A. (1975). Cross-lagged panel correlation: A test for spuriousness. *Psychological Bulletin, 82,* 887–903. http://dx.doi.org/10.1037/0033-2909.82.6.887

Koponen, T., Salmi, P., Ekland, K., & Aro, T. (2013). Counting and RAN: Predictors of arithmetic calculation and reading fluency. *Journal of Educational Psychology, 105,* 162–175. http://dx.doi.org/10.1037/a0029285

Lee, K., & Bull, R. (2016). Developmental changes in working memory, updating, and math achievement. *Journal of Educational Psychology, 108,* 869–882. http://dx.doi.org/10.1037/edu0000090

LeFevre, J.-A., Fast, L., Skwarchuk, S.-L., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. *Child Development, 81,* 1753–1767. http://dx.doi.org/10.1111/j.1467-8624.2010.01508.x

Li, Y., & Geary, D. C. (2013). Developmental gains in visuospatial memory predict gains in mathematics achievement. *PLoS One, 8*(7), e70160. http://dx.doi.org/10.1371/journal.pone.0070160

Liu, Z., & Guo, W. (2011). fmixed: A SAS macro for smoothing-spline-based functional mixed effects models. *Journal of Statistical Software, 43*(c01). http://dx.doi.org/10.18637/jss.v043.c01

Ma, X. (1999). A meta-analysis of the relationship between anxiety toward mathematics and achievement in mathematics. *Journal for Research in Mathematics Education, 30,* 520–540. http://dx.doi.org/10.2307/749772

Mazzocco, M. M., & Kover, S. T. (2007). A longitudinal assessment of executive function skills and their association with math performance. *Child Neuropsychology, 13,* 18–45. http://dx.doi.org/10.1080/09297040600611346

Moore, A. M., vanMarle, K., & Geary, D. C. (2016). Kindergartners' fluent processing of symbolic numerical magnitude is predicted by their cardinal knowledge and implicit understanding of arithmetic 2 years earlier. *Journal of Experimental Child Psychology, 150,* 31–47. http://dx.doi.org/10.1016/j.jecp.2016.05.003

Müller, H. G. (2009). Functional modeling of longitudinal data. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 223–252). New York, NY: Wiley.

Östergren, R., & Träff, U. (2013). Early number knowledge and cognitive ability affect early arithmetic ability. *Journal of Experimental Child Psychology, 115,* 405–421. http://dx.doi.org/10.1016/j.jecp.2013.03.007

Pickering, S., & Gathercole, S. (2001). *Working Memory Test Battery for Children (WMTB-C) manual.* London, England: Psychological Corporation Ltd.

Preacher, K. J. (2015). Advances in mediation analysis: A survey and synthesis of new developments. *Annual Review of Psychology, 66,* 825–852. http://dx.doi.org/10.1146/annurev-psych-010814-015258

Qin, S., Cho, S., Chen, T., Rosenberg-Lee, M., Geary, D. C., & Menon, V. (2014). Hippocampal-neocortical functional reorganization underlies children's cognitive development. *Nature Neuroscience, 17,* 1263–1269. http://dx.doi.org/10.1038/nn.3788

Ramsay, J. O., Hooker, G., & Graves, S. (2009). *Functional data analysis with R and Matlab.* New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-98185-7

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135,* 943–973. http://dx.doi.org/10.1037/a0017327

Rivera-Batiz, F. (1992). Quantitative literacy and the likelihood of employment among young adults in the United States. *The Journal of Human Resources, 27,* 313–328. http://dx.doi.org/10.2307/145737

Rogosa, D. (1980). A critique of cross-lagged correlation. *Psychological Bulletin, 88,* 245–258. http://dx.doi.org/10.1037/0033-2909.88.2.245

Rouder, J. N., & Geary, D. C. (2014). Children's cognitive representation of the mathematical number line. *Developmental Science, 17,* 525–536. http://dx.doi.org/10.1111/desc.12166

Sameroff, A. J., Seifer, R., Baldwin, A., & Baldwin, C. (1993). Stability of intelligence from preschool to adolescence: The influence of social and family risk factors. *Child Development, 64,* 80–97. http://dx.doi.org/10.2307/1131438

SAS Institute. (2014). *Statistical analysis system 9.2.* Cary, NC: Author.

Schmidt, F. L., & Crano, W. D. (1974). A test of the theory of fluid and crystallized intelligence in middle-and low-socioeconomic-status children: A cross-lagged panel analysis. *Journal of Educational Psychology, 66,* 255–261. http://dx.doi.org/10.1037/h0036093

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116,* 250–264. http://dx.doi.org/10.1037/0096-3445.116.3.250

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child Development, 75,* 428–444. http://dx.doi.org/10.1111/j.1467-8624.2004.00684.x

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., . . . Chen, M. (2012). Early predictors of high school mathematics achievement. *Psychological Science, 23,* 691–697. http://dx.doi.org/10.1177/0956797612440101

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology, 62,* 273–296. http://dx.doi.org/10.1016/j.cogpsych.2011.03.001

Sweller, J. (2012). Human cognitive architecture: Why some instructional procedures work and others do not. In K. R. Harris, S. Graham, & T. Urdan, (Eds.), *Educational psychology handbook. Vol 1: Theories, constructs, and critical issues* (pp. 295–325). Washington, DC: American Psychological Association.

Thorndike, R. L. (1933). The effect of the interval between test and retest on the constancy of the IQ. *Journal of Educational Psychology, 24,* 543–549. http://dx.doi.org/10.1037/h0070255

Thorsen, C., Gustafsson, J. E., & Cliffordson, C. (2014). The influence of fluid and crystallized intelligence on the development of knowledge and skills. *The British Journal of Educational Psychology, 84,* 556–570. http://dx.doi.org/10.1111/bjep.12041

Tricot, A., & Sweller, J. (2013). Domain-specific knowledge and why teaching generic skills does not work. *Educational Psychology Review, 26,* 265–283. http://dx.doi.org/10.1007/s10648-013-9243-1

Van de Weijer-Bergsma, E., Kroesbergen, E. H., & Van Luit, J. E. (2015). Verbal and visual-spatial working memory and mathematical ability in different domains throughout primary school. *Memory & Cognition, 43,* 367–378. http://dx.doi.org/10.3758/s13421-014-0480-4

von Aster, M. G., & Shalev, R. S. (2007). Number development and developmental dyscalculia. *Developmental Medicine and Child Neurology, 49,* 868–873. http://dx.doi.org/10.1111/j.1469-8749.2007.00868.x

Wang, J.-L., Chiou, J.-M., & Müller, H.-J. (2016). Functional data analysis. *Annual Review of Statistics and Its Application, 3,* 257–295. http://dx.doi.org/10.1146/annurev-statistics-041715-033624

Watts, T. W., Duncan, G. J., Chen, M., Claessens, A., Davis-Kean, P. E., Duckworth, K., . . . Susperreguy, M. I. (2015). The role of mediators in the development of longitudinal mathematics achievement associations. *Child Development, 86,* 1892–1907. http://dx.doi.org/10.1111/cdev.12416

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher, 43,* 352–360. http://dx.doi.org/10.3102/0013189X14553660

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence.* San Antonio, TX: Psychological Corporation.

Wechsler, D. (2001). *Wechsler Individual Achievement Test–II: Abbreviated.* San Antonio, TX: The Psychological Corporation.

Received June 10, 2016
Revision received August 16, 2016
Accepted September 1, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 5, 694–708

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000169

Working Memory Strategies During Rational Number Magnitude Processing

Michelle Hurst and Sara Cordes
Boston College

Rational number understanding is a critical building block for success in more advanced mathematics; however, how rational number magnitudes are conceptualized is not fully understood. In the current study, we used a dual-task working memory (WM) interference paradigm to investigate the dominant type of strategy (i.e., requiring verbal WM resources vs. requiring primarily visuospatial WM resources) used by adults when processing rational number magnitudes presented in both decimal and fraction notation. Analyses revealed no significant differences in involvement of verbal and visuospatial WM, regardless of notation (fractions vs. decimals), indicating that adults rely upon a mix of strategies and WM resources when processing rational number magnitudes. However, this pattern interacted with algebra ability such that those performing better on the algebra assessment relied upon both verbal and visuospatial WM when engaging in rational number comparisons, whereas rational number performance by adults with low algebra fluency was affected only by a simultaneous verbal WM task. Together, results support previous work implicating the involvement of WM resources in rational number processing and is the first study to indicate that the involvement of both verbal and visuospatial WM, as opposed to relying primarily on verbal WM, when processing rational number magnitudes may be indicative of higher mathematical proficiency in the domain of algebra.

Keywords: working memory, rational numbers, fractions, decimals, algebra

An understanding of rational number concepts has been shown to be critical for further math learning. For example, early fraction and decimal knowledge is a unique predictor of arithmetic ability and general math achievement in elementary and middle school (e.g., Bailey, Hoard, Nugent, & Geary, 2012; Schneider, Grabner, & Paetsch, 2009), as well as algebra ability in older children and adults (e.g., Booth, Newton, & Twiss-Garrity, 2014; Hurst & Cordes, 2016b; Siegler et al., 2012). Although substantial evidence suggests that there is some relationship between algebra and rational number ability, what aspect of rational number knowledge is most critical and the mechanisms through which this relationship develops are only just beginning to be explored. Recent evidence has suggested that one of the critical aspects of fraction knowledge is an understanding of rational number magnitudes (Booth & Newton, 2012; Booth et al., 2014). For example, Booth et al. (2014) found that eighth graders' ability to map fractions onto number lines was predictive of improvement in their equation solving after an algebra course. However, processing rational number magnitudes is not a straightforward task, as it potentially involves distinct strategies, ranging from holistic processing (i.e., getting a feel for the numerical size of a value) to computational processing (i.e., transforming fractions into decimal values to get a sense for the size of the value). Yet no work has investigated whether specific rational number magnitude processing strategies may be stronger predictors of algebraic processing. Thus, in order to better understand the relationship between rational number magnitude understanding and algebra ability, we must also investigate how people go about processing rational number magnitude information and whether there are differences in how these magnitudes are understood across individuals with differing algebra abilities.

Rational Number Magnitudes

To investigate how people think about the magnitudes associated with symbolic numbers, researchers often use number comparison tasks. In these tasks, participants are asked to rapidly judge which of two numbers is greater. Work with whole numbers has revealed that performance on these tasks is predictive of more general math fluency (e.g., Holloway & Ansari, 2009) and correlated with math anxiety (e.g., Maloney, Ansari, & Fugelsang, 2011), suggesting that performance on these tasks can provide insight into how these values are processed. More recently, researchers have begun to use these tasks with other types of numbers like fractions and decimals, with results similarly revealing performance on these rational number magnitude comparisons predicting math ability in other domains (e.g., Hurst & Cordes, 2016b; Schneider et al., 2009; Siegler, Thompson, & Schneider, 2011).

However, little is known about the specific strategies children and adults may invoke to access the magnitudes associated with

This article was published Online First January 2, 2017.

Michelle Hurst and Sara Cordes, Department of Psychology, Boston College.

Funding provided by the Natural Sciences and Engineering Research Council of Canada to Michelle Hurst.

Correspondence concerning this article should be addressed to Michelle Hurst, Department of Psychology, Boston College, 140 Commonwealth Avenue, 300 McGuinn Hall, Chestnut Hill, MA 02467. E-mail: hurstm@bc.edu

rational number notation. Some evidence suggests that in a rational number comparison task (e.g., "Which is larger [1/2] vs. [3/4]?"), adults are able to access magnitude information from both fractions and decimals (e.g., DeWolf, Grounds, Bassok, & Holyoak, 2014; Hurst & Cordes, 2016a; Schneider & Siegler, 2010), but only when they are prevented from using other component-based strategies (Bonato, Fabbri, Umiltà, & Zorzi, 2007). When explicitly asked to report their strategy use, Faulkenberry and Pierce (2011) found that adults' strategies could be grouped into one of five different categories (although a small percentage reported strategies that did not fit into one of these categories): just knowing it, cross-multiplication, benchmarking (e.g., comparing the values with [1/2]), visualization, and converting fractions into decimals. Although many of these strategies involve understanding magnitude, they may also involve other procedures—including arithmetic and calculation. Given that fractions and decimals are complicated symbols that involve a combination of Arabic numerals and non-numeric symbols (i.e., the vinculum, or dividing line, in fractions and the decimal point in decimals), it may not be surprising that some adults engage in calculation-based strategies (i.e., cross-multiplication, converting fractions into decimals, and possibly benchmarking). In addition, even the strategy of "just knowing it" (the single strategy with the highest reported use—30.7% of trials; Faulkenberry & Pierce, 2011) may have encompassed more than one type of implicit strategy, including ones the participants could not readily describe using self-report. Thus, given that fraction and decimal magnitudes may be interpreted and processed in different ways, and that fraction and decimal magnitude understanding is related to algebra ability, it is important to explore whether differences in how rational number magnitudes are approached may be related to algebra ability. However, because self-report may not be the most accurate way to assess strategies, and because the reporting of strategies on each trial could potentially impact the future use of those strategies within the task, it is ideal to investigate rational number magnitude strategies using implicit measures.

Working Memory

To implicitly assess rational number processing strategies, the current study explored how distinct components of WM (i.e., the phonological loop [i.e., verbal WM] and the visuospatial sketchpad; Baddeley, 1992; Baddeley, 2012; Baddeley & Hitch, 1974) are implicated during a rational number magnitude task. Although studies have identified a relationship between WM capacity and rational number processing abilities (e.g., Jordan et al., 2013; Vukovic et al., 2014), no studies have explored how these distinct components of WM may individually contribute to rational number magnitude processing. Importantly, understanding the involvement of these distinct components of WM during mathematical tasks can provide insight into the nature of the strategies invoked when performing these tasks (e.g., Caviola, Mammarella, Cornoldi, & Lucangeli, 2012; DeStefano & LeFevre, 2004; Raghubar, Barnes, & Hecht, 2010).

Significant research has explored how these WM components are implicated in other numerical and math tasks, such as mental arithmetic, revealing distinct patterns of involvement for the phonological loop and the visuospatial sketchpad. Given that the phonological loop is thought to be involved in temporarily storing verbal information in memory (Baddeley, 1992, 2012; Baddeley & Hitch, 1974), it is not surprising that the phonological loop is implicated in mental arithmetic tasks in which verbal strategies are invoked, such as when children use counting strategies and/or perform calculations that involve maintaining operands or an interim solution (DeStefano & LeFevre, 2004). Even in adults, verbal WM has been shown to be involved in solving complex arithmetic problems (Hitch, 1978). The visuospatial sketchpad, on the other hand, is deemed responsible for maintaining visual information in memory, including creating mental pictures and diagrams (Baddeley, 1992, 2012; Baddeley & Hitch, 1974). Thus, the visuospatial sketchpad has been found to play a role when mental transformation of the problem may be necessary for solving the problem (e.g., carrying in multidigit addition presented vertically; Caviola et al., 2012). In addition, theories positing that individuals use a "mental blackboard" to solve mathematical problems (e.g., Hayes, 1972) suggest that the visuospatial sketchpad may be involved to some extent in most situations of arithmetic, although the specific role of the visuospatial sketchpad in mental arithmetic, and particularly complex arithmetic, is unclear (e.g., Hubber, Gilmore, & Cragg, 2014).

Other work reveals that numerical magnitudes themselves, specifically for whole numbers, are visuospatially encoded in both adults and children (Simmons, Willis, & Adams, 2012; van Dijck, Gevers, & Fias, 2009). This is consistent with findings from other tasks suggesting that both children and adults represent whole numbers along a spatially encoded mental number line (e.g., Dehaene, Bossini, & Giraux, 1993; Moyer & Landauer, 1967, 1973), which may suggest that magnitude processing—distinct from mental arithmetic—may rely primarily on visuospatial WM and only minimally involve verbal WM. Whether this is also the case for rational numbers—whose magnitudes can be accessed through visualization (i.e., envisioning a pie chart), through verbally based strategies (e.g., direct computation, such as converting a fraction to a decimal or step-by-step digit comparisons in decimals), or a mix of strategies (e.g., estimating on a number line using place value or spatially demanding computations, such as cross-multiplication)—is an open question.

To investigate how distinct WM components may be implicated in rational number processing, in the current study, we employed a dual-task WM paradigm, which involves performing a primary task of interest (e.g., number comparison) while performing a secondary task that intentionally taxes WM resources (e.g., remembering four letters). The dual-task WM paradigm has been used to investigate how these various components of WM may be implicated during the primary task in order to identify implicit strategies involved (e.g., DeStefano & LeFevre, 2004; Raghubar et al., 2010). Importantly, the assumption of dual-task paradigms is that if processing in both the primary and the secondary tasks rely upon the same cognitive resources (e.g., verbal WM), then performance on the primary task will be impaired in the dual-task paradigm relative to that of a single-task control. On the other hand, if both tasks can be performed simultaneously without any interference, then they must not rely upon the same cognitive resources.

Given that symbolic notation for both fraction and decimal magnitudes involve both numeric (i.e., Arabic numerals) and non-numeric (i.e., decimal point, vinculum/division bar) symbols, there is reason to expect the involvement of both visuospatial and verbal

resources when adults conceive of rational number magnitudes. On the one hand, given the rampant use of visual representations in the classroom when teaching rational numbers, coupled with evidence suggesting that fractions and decimals are spatially encoded (Faulkenberry & Pierce, 2011; Hurst & Cordes, 2016a; Schneider & Siegler, 2010), values in fraction and decimal notation may be processed holistically as magnitudes, without engaging explicit computations (e.g., envisioning a pie chart and/or as values falling along a line). If so, then rational number magnitudes should be primarily visuospatially encoded, only minimally requiring the use of the phonological loop. Alternatively, fraction notation implies the division of two whole numbers and decimals involve multiple components (i.e., values before and after the decimal point), making interpretation of the magnitudes associated with these symbols less transparent. In turn, this might suggest that both fraction and decimal magnitudes may be only accessible via direct computation and/or component-to-component comparisons (e.g., cross-multiplication, comparing values in the tenths digit, then in the hundredths, and so on). If this is the case, then processing of these values should require a greater reliance upon the phonological loop to maintain interim solutions in the calculation for each comparison.

Notation Differences

Furthermore, although fraction and decimal notation are used to represent the same numerical magnitudes, processing of values in these distinct notations may not rely on the same WM resources. It has been argued that decimal notation is more similar to whole numbers (e.g., Johnson, 1956), and recent evidence suggests that magnitudes are more easily accessed in decimal notation relative to fraction notation (e.g., DeWolf et al., 2014; Hurst & Cordes, 2016a). If so, then judgments of decimal magnitudes (as opposed to fraction magnitudes) may be more likely to be spatially encoded (similar to whole-number magnitudes), and thus rely primarily upon visuospatial WM, whereas fraction magnitudes may reveal a greater reliance upon verbal resources (reflecting increased computations, i.e., translating into decimal notation). Alternatively, evidence suggests that adults conceive of fraction and decimal magnitudes as falling along a single integrated mental continuum (Hurst & Cordes, 2016a), suggesting that underlying similarities in the numerical concepts these distinct notations represent may be salient to adults. If so, then these symbolic systems may receive similar treatment, resulting in consistent strategies employed across notations.

In addition, if distinct strategies are employed when comparing magnitudes exclusively in decimal notation (by comparing two decimals; e.g., 0.5 vs. 0.75) and exclusively in fraction notation (by comparing two fractions; e.g., [1/2] vs. [3/4]), then investigating the strategies used when comparing two values presented in different notation (by comparing a fraction with a decimal; e.g., [1/2] vs. 0.75) can provide important insight. For example, if the level of verbal WM recruitment increases as a function of the number of fractions involved in the comparison (with the lowest level of recruitment involved in comparisons between two decimals [zero fractions], with slightly more for comparisons between a decimal and a fraction [one fraction], and the highest level for those between two fractions), then this would suggest that each fraction requires additional computational processing. Alternatively, if comparisons involving two fractions yield the same pattern as comparisons involving a decimal and a fraction, then it may be that merely the presence of a fraction invokes a distinct set of strategies not employed when there are only decimals.

Relationship to Algebra Ability

Most critically, however, is investigating whether the involvement of visuospatial and verbal WM resources may differ as a function of algebra ability. The relationship between algebra ability and rational number understanding has been explained through a number of mechanisms, including having a strong understanding of the rational number system, being proficient with both algebraic and arithmetic procedures, understanding the conceptual aspects of fraction units (e.g., the denominator), and so on (e.g., Booth & Newton, 2012; Hurst & Cordes, 2016b; Kilpatrick & Izsak, 2008; Wu, 2001). Although studies have investigated algebra (e.g., Booth & Davenport, 2013; Koedinger, Alibali, & Nathan, 2008; Landy, Brookes, & Smout, 2014) and fraction problem solving (e.g., Faulkenberry & Pierce, 2011) separately, how specific strategies for approaching fraction problems may be related to algebra proficiency is an unexplored area. Given that understanding fraction magnitudes may be critical for algebra understanding (e.g., Booth & Newton, 2012; Kilpatrick & Izsak, 2008), we might expect those proficient in algebra to engage in fewer computational strategies when processing fraction and decimal magnitudes (having a more intuitive understanding of the magnitudes associated with those symbols) and those less proficient in algebra to rely more upon calculation-based strategies. If this is the case, then we would expect to see individuals with lower algebra fluency to rely more upon verbal WM resources and less so upon visuospatial WM resources. On the other hand, other evidence suggests that performance on rational number arithmetic assessments is also predictive of algebra ability and may be an essential part of the relationship (Hurst & Cordes, 2016b; Kilpatrick & Izsak, 2008). Thus, it may be that those who are more fluent with algebraic processing are more likely to process fraction and decimal magnitudes arithmetically, executing calculations in order to make the comparison, for example, cross-multiplying two fractions or converting values into a common notation for purposes of comparison. In this case, we might expect that those individuals with higher algebra ability to have greater reliance upon verbal WM resources (evidence of engaging a calculation based strategy), whereas those with lower algebra ability may not.

The Current Study

In summary, there is a growing literature investigating how people think about rational number magnitudes and how rational number understanding may be related to algebra ability. However, there are several open questions about the strategies invoked when processing rational number magnitudes presented in both fraction and decimal notation. In the current study, we used a dual-task WM paradigm to assess how visuospatial and verbal WM are implicated during a rational number magnitude comparison task. We then assessed whether individual differences in WM involvement (indicative of distinct processing strategies) were associated with performance on an algebraic assessment. We explored these relationships in a group of educated young adults who have had several years of schooling beyond the introduction of basic rational

number and algebra concepts. Given that rational number and algebra concepts are introduced in different school grades and taught throughout a large range of grades (Common Core State Standards: National Governors Association Center for Best Practices & Council of Chief of State School Officers, 2010), adult participants allow us to investigate these relationships once they have already received basic educational instruction on these topics. By doing so, we are able to take a first look at the pattern of these relationships, providing insight into individual differences in rational number and algebra understanding and opening up new avenues for further investigation into children who are in the process of learning these concepts.

Specifically, this study addresses three research questions (RQs):

1. Is WM differentially implicated in rational number magnitude understanding based on WM type (visuospatial vs. verbal WM)?

RQ #1 will be investigated by looking at whether performance on the rational number task differs depending on WM load type. If rational number magnitudes are processed in terms of visuospatial representations (i.e., visualizing the quantities), then we would expect visuospatial WM to show more interference than verbal WM. On the other hand, if rational number magnitudes are primarily processed in terms of their computational features (i.e., arithmetic manipulation of the symbols), then we would expect primarily verbal WM interference.

2. Is WM differentially implicated in rational number magnitude understanding based on rational number notation (fractions vs. decimals)?

RQ #2 will be investigated by looking at whether the level of verbal and/or visuospatial WM interference depends upon the notation being used. If fractions and decimals are processed similarly, we would expect no differences across notation. Alternatively, given substantial literature suggesting adults consider these notations to be qualitatively different (e.g., DeWolf et al., 2014; Hurst & Cordes, 2016a), we might expect WM interference to differentially impact decimals and fractions.

3. Do individuals with different levels of algebra ability show distinct patterns of WM resource use in a rational number magnitude task? That is, does the pattern of findings in RQ #1 depend on the algebra ability of the individual?

RQ #3 will be investigated by looking at how the pattern of results discussed in RQ #1 may differ across those with high and low algebra ability. If the often-reported relationship between algebra ability and rational number understanding (e.g., Booth et al., 2014; Hurst & Cordes, 2016b; Siegler et al., 2012) is dependent upon the type of resource-based strategies used by the individual, then we would expect to see different levels of verbal and visuospatial WM involvement between those with high and low algebra ability. Furthermore, in order to isolate algebra ability in particular, we will include performance on a rational number arithmetic assessment as a covariate in our analyses in order to control for individual differences in procedural ability with rational number notation more generally.

By investigating individual differences in WM recruitment, we may be able to look at differences in how those with varying algebra abilities approach rational number magnitudes. By relying on previous research with WM recruitment during mental arithmetic, we may be able to shed some light on the kinds of strategies adults may employ based on their patterns of WM recruitment.

Method

Participants

Seventy-nine adults participated for course credit or $10.00. Adults were recruited from a university campus through introductory psychology courses and flyers, resulting in a sample primarily made up of undergraduate and graduate college students. Nineteen adults were not included in the analyses because of computer error resulting in the loss of all data ($n = 13$) or because their data exceeded our exclusion criteria ($n = 6$; see the Exclusion Criteria section for details). Thus, data from a final sample of 60 adults ($M_{age} = 20.9$ years; age range = 18 to 33 years old; 35% males) were included in the WM analyses. Additionally, data from three adults were excluded from analyses involving the math assessments (see Exclusion Criteria), resulting in data from 57 adults ($M_{age} = 21.0$ years; age range = 18 to 33 years old; 37% males) used for analyses relating WM involvement to algebra performance.

Procedure

All adults completed the number comparison task, participating in four within-subject blocks (visuospatial dual-task, visuospatial control, verbal dual-task, verbal control), with the order of the blocks counterbalanced across subjects. Each individual block of the WM dual-task took approximately 5 min. Following the number comparison task, adults completed two math assessments: a rational number arithmetic assessment and an algebraic assessment. The entire session took no longer than 60 min.

Each block of the WM dual-task began with three practice problems (one of each number comparison type), during which the experimenter sat next to the participant to ensure that the participant understood and followed the instructions. During the task, the experimenter left the room and only reentered to provide instructions for the next condition. All tasks (number comparison task and math assessments) were presented on a 22-in. monitor connected to an Apple computer.

Measures

WM dual-task. The WM dual-task procedure (modeled after Caviola et al., 2012) contained four within-subject blocks: two dual-task blocks (visuospatial and verbal) and two control blocks (visuospatial and verbal). On the dual-task blocks, participants were asked to remember visuospatial or verbal information (secondary task) while performing a numerical comparison (primary task). The control blocks were designed to be perceptually and temporally identical to the dual-task blocks (i.e., to have the same temporal spacing between trials and the same perceptual distrac-

tors)—the only difference was that participants were instructed not to remember the information from the secondary task, and they were never asked to recall that information. Trials in each block followed the same basic procedure: (a) center fixation cross (1,000 ms; 1.5 cm × 1.5 cm); (b) secondary task memory stimulus (2,500 ms); (c) blank screen (1,000 ms); (d) number comparison stimuli (until response); (e) blank screen (1,000 ms); and (f) memory recall (in dual blocks) or memory stimulus reappearance (in control blocks; until the participant responded to move on to next trial; see Figure 1).

Primary task. The primary task of interest was the number comparison task (similar to tasks used in DeWolf et al., 2014; Hurst & Cordes, 2016a). Across all four blocks, participants were presented two rational numbers and were instructed to indicate which of the two numbers was larger in numerical value. There were three different types of numerical comparison trials in which participants were asked to judge the relative magnitude of: (a) two fractions (FvF; e.g., [1/2] vs. [3/4]); (b) two decimals (DvD; e.g., 0.5 vs. 0.75); or (c) one decimal and one fraction (DvF; e.g., 0.5 vs. [3/4]). Participants indicated their response by selecting the corresponding key (right arrow for right stimulus and left arrow for left stimulus) on the keyboard as quickly as possible.

In each of the four blocks, eight trials of each number comparison type (FvF, DvD, and DvF; all intermixed) were randomly presented, for a total of 96 trials (8 trials × 3 comparison types × 4 blocks).

On the FvF and DvF trials, the two numerical values presented differed, on average, by a ratio of 2.4 (range = 2.1 to 2.8), and the numerical values presented in DvD comparisons differed by an average ratio of 1.12 (range = 1.07 to 1.15). The ratio of the DvD comparisons was set lower than the DvF and FvF comparisons, because previous work (Hurst & Cordes, 2016a) suggested that a ratio around 1.12 in DvD comparisons would result in a comparable level of performance as DvF and FvF trials at a 2.4 ratio.

Numerators and denominators of the fraction stimuli only involved single-digit values ranging from 1 to 9 and had magnitudes between 0 and 2 (exact range = 1/7 to 9/5), resulting in a mix of proper unit fractions, proper nonunit fractions, and improper fractions. In FvF trials, the two numerators and two denominators used to make up the two fractions were always four distinct integers to prevent adults from using an exclusively numerator or denominator comparison strategy (as in Schneider & Siegler, 2010).

Decimal values ranged from 0.15 to 1.69 (approximately the same range as the fractions), making the unit value in the decimal notation (i.e., value to the left of the decimal point) either a 0 or a 1. Every DvD trial included one numerical stimulus with digits to the thousandths place (i.e., had three digits after the decimal; e.g., 0.714) and the other included digits only to the hundredths place (i.e., two digits after the decimal; e.g., 1.75). On half the trials, the correct (larger value) decimal was the longer decimal (decimal with three digits), and on the other half, the correct decimal value was the shorter decimal (decimal with two digits), in order to make decimal length (i.e., number of digits after the decimal point) not a reliable indicator of which magnitude was larger. The Appendix provides the full set of numerical stimuli.

All stimuli were made in 100-point Myriad Pro font (in Adobe Illustrator) and were approximately 5 cm apart (from right edge of the left stimulus to the left edge of the right stimulus), centered on the screen. Fraction stimuli were approximately 3.5 cm wide × 5 cm high; decimal stimuli to the thousandths place were approximately 5.5 cm wide × 2 cm high; and decimal stimuli to the hundredths place were approximately 4.25 cm wide × 2 cm high.

Reaction time (RT) was used as the primary dependent variable. Only RTs from correct trials and those within three standard deviations of the individual participant's average RT on that notation and WM condition were included.

Secondary tasks. During the dual-task blocks, participants engaged in a secondary task at the same time as performing the primary task.

Secondary visuospatial task. In the visuospatial dual-task block, participants performed a secondary visuospatial task (while performing the primary task) in order to tax visuospatial memory

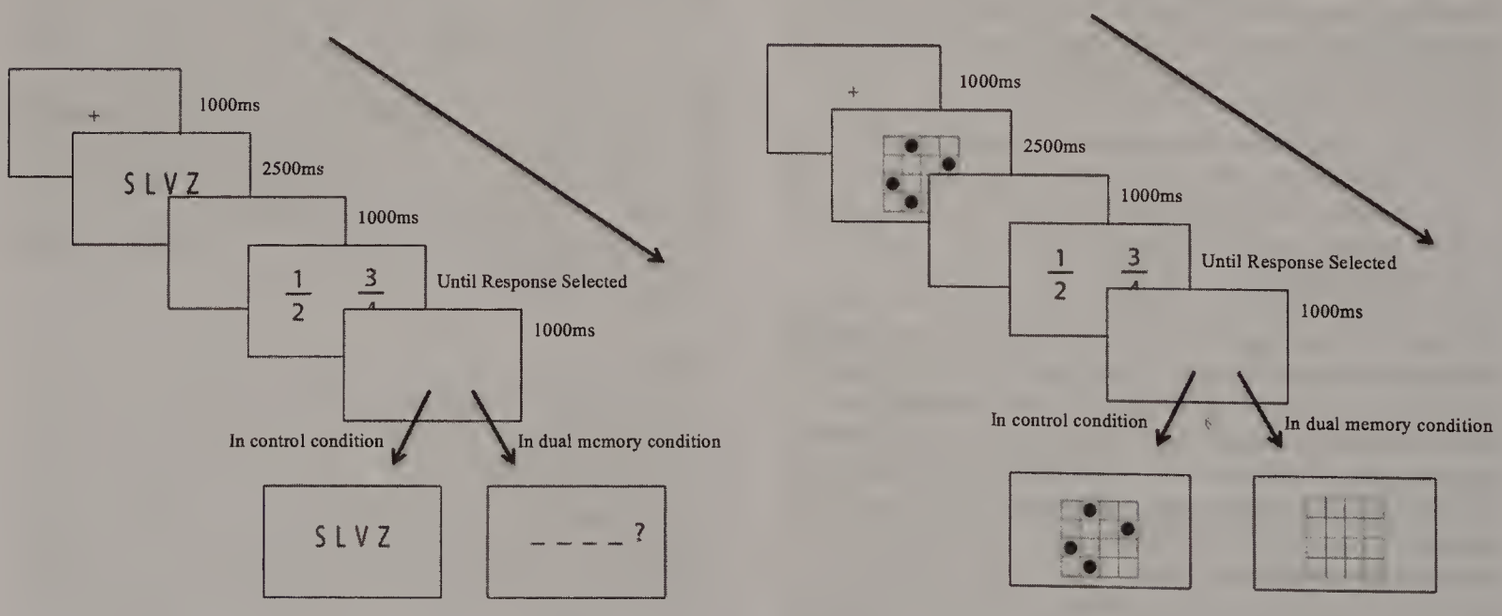


Figure 1. The procedure for the verbal working memory (WM) conditions (left) and the visuospatial WM conditions (right).

resources. The secondary visuospatial task required participants to remember visuospatial information on every trial. Participants were presented with a 240.25-cm^2 4 × 4 grid (made up of 16 3.75 cm × 3.75 cm squares). The grid was centered on the screen and four of the locations on the grid contained a black circle (2.75 cm in diameter, centered within the square on the grid).

In the dual-task block, on every trial, participants were instructed to remember the location of the four circles shown on the grid while performing a numerical comparison. Although participants were told to remember the information on every trial, they were only asked to recall this information on a random half of the trials (in order to shorten the experiment length; as in Caviola et al., 2012). On trials in which participants were asked to recall the visuospatial information, after selecting their response to the numerical comparison, participants were shown an empty 4 × 4 grid and were instructed to click on the four locations in the grid (using the computer mouse) in which they recalled there being a black circle. After selecting the four locations, they pressed the up-arrow key on the keyboard to submit their response and move on to the next trial. On those trials in which they were not asked to recall the location of the circles, participants were shown the same 4 × 4 grid (including the four circles) they had seen prior to the number comparison. On these trials, they simply had to press the up-arrow to move on to the next trial. Whether the participant had to recall the grid stimulus or not was randomly determined on a trial-to-trial basis, such that participants could not reliably pick and choose when to remember the information and when not to remember. Thus, in order to succeed in the task, they were required to remember the grid on every trial. Accuracy on those trials in which subjects were asked to recall the visuospatial information (in dual blocks) was scored to ensure that participants actually remembered the information during the dual-task conditions. Visuospatial WM accuracy was measured as the number of trials in which the participant indicated the correct location of at least three of the four dots in the grid.

In the visuospatial control block, trials were identical to the visuospatial dual-task block, except that participants were never asked to recall the locations of the circles. That is, on every trial, participants were shown a 4 × 4 grid with four circles before the number comparison task, and were reshown the same 4 × 4 grid after the number comparison task, and had to push the up-arrow to move on to the next trial. Importantly, participants were told that they did not need to remember the location of the circles on the grid, thus making it irrelevant to the task (though identical to the dual-task block in every other way).

None of the visuospatial memory stimuli were presented more than once to each person, resulting in 48 (plus six practice) different visuospatial stimuli. However, the same 24 numerical comparison stimuli were used for both of the visuospatial blocks (but were different than the verbal block numerical stimuli), and all the visuospatial stimuli were randomly paired with their accompanying number comparison stimuli for each participant.

Secondary verbal task. In the verbal dual-task block, participants performed a secondary verbal task (while performing the primary numerical comparison task) in order to tax verbal memory resources. The secondary verbal task involved remembering verbal information (as in Caviola et al., 2012). Participants were presented with four consonant letters from the English alphabet in a random order centered on the screen (e.g., XQRT). The total length of the four letters was approximately 12 cm × 3 cm.

The verbal dual-task block was identical to the visuospatial dual-task block, except that on every trial, instead of a grid, participants were shown four consonant letters (presented horizontally) and were instructed to read them out loud and remember them. Following the number comparison task, on half the trials, participants were provided with an empty text box and instructed to type in the four letters they saw previously, then press the up-arrow to submit their response and move on to the next trial. On the other half of the trials, participants were not asked to recall the verbal information, but instead were reshown the same four letters and just had to press the up-arrow to move on to the next trial. Accuracy on those trials in which subjects were asked to recall the verbal information was scored to ensure that participants remembered the required information during the dual-task blocks. Verbal WM accuracy was scored as the number of trials in which the participant correctly recalled all four letters.[1]

The verbal control block was perceptually and temporally identical to the half of the trials in the verbal dual-task block that did not require the participant to recall the letters they had seen previously. Thus, participants were instructed to read the four letters aloud but not to remember them. After the number comparison task, participants were reshown the same four letters and simply had to press the up-arrow to move on to the next trial.

None of the verbal memory stimuli were presented more than once to each person, resulting in 48 (plus six practice) different verbal stimuli. However, the same 24 numerical comparison stimuli were used for both of the verbal blocks (but were different than the visuospatial block numerical stimuli), and all of the verbal stimuli were randomly paired with the accompanying number comparison stimuli for each participant.

Math assessments. Following the dual-task procedure, participants completed math assessments given in two parts: rational number arithmetic (involving both fractions and decimals) and algebra, in that order. For all assessments, questions were presented one at a time on a computer screen, and participants were given a paper workbook to do as much work as they needed and to record their answers. The use of aid devices (e.g., calculators) was not allowed. Participants were told they had as much time as they needed, but to work as quickly as they could because they were being timed (by the computer).

The fraction and decimal assessment consisted of eight decimal arithmetic problems and eight fraction arithmetic problems, presented in two blocks with order counterbalanced (see the Appendix for a full list of problems). There were two each of addition, subtraction, division, and multiplication problems for each notation type. The fraction problems always contained four distinct integers, meaning none of the problems contained a common denominator or a common numerator. For the decimal problems, one problem of each arithmetic type involved two decimal values with the same number of digits (i.e., to the hundredths digit, e.g., 0.48 + 0.56). The other problem within each arithmetic type

[1] Slightly different criteria were used for the visuospatial and verbal blocks in order to approximately match accuracy between visuospatial (average 91% correct as opposed to 79% when the same criteria were used) and verbal (average 94% correct) working memory.

involved two decimal values with a different number of digits (e.g., 0.67 + 0.843).

The algebra assessment consisted of 12 problems adapted from the Trends in International Mathematics and Science Study (TIMSS) Grade Eight assessment (International Association for the Evaluation of Educational Achievement [IEA], 2005, 2013). The assessment involved a variety of problems involving solving expressions, using values in an expression, and finding the relation between values in a table or expressed in a word problem (see the Appendix for a full list of problems). Importantly, correctly solving the algebra problems only required manipulation of whole numbers (noninteger values were not included in this assessment). Thus, although whole-number division was occasionally required (e.g., 24/8 = 3), no knowledge of arithmetic or procedures associated with fractions and decimals was required.

Two independent coders scored each of the math assessments to determine accuracy (99% agreement on both assessments, with a third coder resolving the disagreements), and the computer recorded completion time. Accuracy was fairly high on both the algebra (M = 9.8 of 12) and rational number arithmetic assessment (M = 12.6 of 16), with relatively low variability (e.g., 50% of adults scored 10, 11, or 12 out of 12 on the algebra measure). Thus, as our dependent measure, we used completion time as a measure of fluency for both the rational number arithmetic and the algebra measures. The internal reliability of the algebra assessment was fairly good, with Cronbach's alpha of 0.744 for this sample (based on completion times).

Exclusion Criteria

WM dual task. As per Caviola et al. (2012; also see Conway et al., 2005), participants were required to score above 60% on both WM types to be included in the analyses in order to ensure they had actually invoked WM during the task. Two participants did not meet this criterion, and so their data were excluded from analyses.

Magnitude comparison task. At the group level, participants who scored below chance on the comparison task (n = 2) or who had RTs greater than three standard deviations away from the group RT performance (n = 2) were excluded. Only RTs from trials in which participants provided the correct response to the numerical comparison were included in analyses. Thus, data from 60 participants were included in analyses of the WM dual-task.

Math assessments. In order to make differences in completion time comparable across individuals (and to avoid issues of speed/accuracy trade-offs), participants who performed worse than 50% correct on the math assessments were excluded from analyses involving the math assessments (n = 3). Thus, data from 57 participants were included in analyses involving the assessments.

Results

Descriptive statistics for each task are presented in Table 1. On all tasks, speed (response time [comparison task] or completion time [assessments]) was used as the primary dependent variable as a measure of fluency. Preliminary analyses suggest no significant main or interaction effects involving gender, and therefore gender was not included in any of the analyses.

Table 1
Means (SDs) for Time and Accuracy for Each Measure

	Verbal WM						Visuospatial WM				Math assessments	
	Control task			Dual task			Control task		Dual task			
	FvF	DvD	DvF	FvF	DvD	DvF	FvF	DvD	DvF	FvF	DvD	DvF	Algebra fluency	Rational number arithmetic fluency
Time	1,277 (295)	1,029 (142)	1,263 (275)	1,491 (525)	1,156 (197)	1,433 (505)	1,252 (473)	1,080 (194)	1,221 (340)	1,333 (387)	1,164 (207)	1,364 (358)	571 (243)	693 (289)
Accuracy	91.7 (8.9)	92.2 (9.5)	95.6 (8.2)	91.7 (8.5)	96.0 (6.7)	94.1 (8.4)	98.3 (5.3)	92.7 (9.8)	95.8 (8.2)	98.1 (6.0)	97.5 (6.4)	97.5 (5.0)	82.5 (12.5)	80.3 (12.4)

Note. Times reported are response times (ms) for the working memory task and completion times (s) for the math assessments. Accuracy is reported as percent correct. WM = working memory; DvD = Decimal vs. Decimal; FvF = Fraction vs. Fraction; DvF = Decimal vs. Fraction.

RQ #1 and RQ #2: Interference Across Notation and WM Type

In order to investigate RQ #1 and RQ #2, we were interested in performance on the dual task to determine whether there were differences in how the secondary WM tasks may have interfered with performance on the primary task depending on the type of WM (RQ #1: visuospatial or verbal) or the notation of the comparison (RQ #2: fractions, decimals, or both). Thus, we used a 2 (task: control vs. dual) × 2 (WM type: visuospatial vs. verbal) × 3 (notation: fraction, decimal, mixed) repeated measures ANOVA on RT.

Consistent with previous research, there was a main effect of task, $F(1, 59) = 38.959$, $p < .001$, $\eta_p^2 = 0.398$, with the dual-tasks ($M_{dual} = 1{,}324$ ms) taking longer than the control tasks ($M_{control} = 1{,}187$ ms). However, task did not interact with WM type ($p = .3$, $\eta_p^2 = 0.02$), suggesting that, in general, there was no evidence of a statistically significant difference in how much the secondary task interfered with RT performance on the primary task across the verbal and visuospatial conditions. Thus, in reference to RQ #1, neither verbal nor visuospatial strategies appeared to dominate over the other. In addition, there was a main effect of notation, $F(1.5, 30.6) = 46.14$, $p < .001$, $\eta_p^2 = 0.44$, with average RT on the DvD trials (1,107 ms) being significantly faster than performance on the FvF (1,338 ms) or DvF (1,320 ms; follow-up t tests, $ps <$ 0.001) trials, which did not significantly differ from each other. However, analyses did not reveal a statistically significant three-way interaction (Type × Notation × WM, $p = .23$, $\eta_p^2 = 0.02$). Thus, in response to RQ #2, there is not statistically significant evidence to suggest that the use of verbal or visuospatial memory differed across the notations.[2]

There was a very small, marginal main effect of WM type, $F(1, 59) = 3.44$, $p = .07$, $\eta_p^2 = 0.06$, with responses in the verbal conditions taking slightly more time ($M_{verbal} = 1{,}275$ ms) than in the visuospatial conditions ($M_{visuo} = 1{,}236$ ms). WM also interacted with notation, $F(1.8, 107.3) = 11.5$, $p < .001$, $\eta_p^2 = 0.16$, such that the difference in response speed between DvD trials and those trials involving fractions was greater in the verbal condition ($M_{DvD} = 1{,}092$ ms, $M_{FvF} = 1{,}384$ ms, $M_{DvF} = 1{,}348$ ms) than in the visuospatial condition ($M_{DvD} = 1{,}122$ ms, $M_{FvF} = 1{,}292$ ms, $M_{DvF} = 1{,}293$ ms). However, none of these variables interacted with task type (Task × Notation, $p = .14$, $\eta_p^2 = 0.03$; Task × Notation × WM, $p = .23$, $\eta_p^2 = 0.02$). Although this pattern may suggest some differences in performance when adults were presented with verbal or visuospatial information during the magnitude comparison task, the critical comparison in our design involved differences between the control and dual-task blocks (i.e., when the verbal/visuospatial information is relevant to the task or not). Therefore, any differences between verbal and visuospatial tasks overall (collapsing across control and dual-task tasks) may be because of perceptual distractions that are not likely related to direct memory effects, making it difficult to interpret these performance differences as meaningful within the current study.

RQ #3: Differences Across Individual Differences in Algebra Fluency

In addition to looking at overall performance on the dual task, we were interested in whether strategy use with rational number magnitudes differed between individuals who were highly proficient in algebra and those who were less proficient, controlling for general math ability. Thus, we divided participants into two groups based on a median split of completion times on the algebra assessment (median completion time = 539 s, range = 252 s to 1,525 s) to create groups that differed in their algebra fluency. We then conducted a 2 (task: control vs. dual) × 2 (WM: visuospatial vs. verbal) × 2 (algebra fluency: high fluency [$N = 28$] vs. low fluency [$N = 29$]) mixed measures ANCOVA on RT on the rational number comparison task, including rational number arithmetic completion time as a covariate (see Figure 2). Because notation did not interact with WM interference in the previous analysis, it was not included as a factor in these analyses. Because completion time on the algebra assessment could assess both general mathematical skills as well as skills specific to algebraic reasoning, completion times on the rational number arithmetic assessment were included as a covariate to allow us to investigate the relationship to algebra ability specifically and not more general math or rational number ability.

In this secondary ANCOVA, which included algebra ability as a factor (and controlled for rational number arithmetic ability), we did not find an overall effect of task ($p = .7$, $\eta_p^2 = 0.003$), WM ($p = .19$, $\eta_p^2 = 0.03$) or an overall Task × WM interaction ($p = .6$, $\eta_p^2 = 0.004$). The rational number arithmetic fluency covariate was related to task type ($p = .04$, $\eta_p^2 = 0.08$), with follow-up analyses suggesting that slower arithmetic performance was positively correlated with overall levels of WM interference (i.e., difference between control and dual-tasks, $r[57] = 0.3$, $p = .03$), indicating that people with higher rational number arithmetic ability were less impacted by the secondary WM task when engaging in rational number comparisons. The rational number arithmetic covariate did not interact with algebra fluency ($ps > 0.1$).

Critically, however, addressing RQ #3, the pattern of WM interference did interact with algebra fluency, suggesting distinct strategy use across those with relatively high versus low algebra fluency. Specifically, there was a WM × Task × Algebra Fluency interaction, $F(1, 54) = 4.76$, $p = .03$, $\eta_p^2 = 0.08$).

A follow up WM × Task ANOVA on higher algebra fluency ($N = 28$) individuals revealed that those with high algebra fluency showed a statistically significant task effect ($p < .001$, $\eta_p^2 = 0.3$, with longer RTs on the dual conditions ($M = 1{,}233$ ms) than on the control conditions ($M = 1{,}109$ ms). However, the task type did not interact with WM type ($p = .16$, $\eta_p^2 = 0.07$), suggesting that data from people with higher algebra fluency showed no evidence of differential involvement of verbal and visuospatial resources, and thus these individuals may have used both verbal and visuospatial strategies approximately equally. On the other hand, when looking at the data from those with lower algebra fluency (2 × 2 ANOVA; $N = 29$), there was a main effect of task ($p < .001$, $\eta_p^2 = 0.43$) as well as a statistically significant Task × WM interaction ($p < .05$, $\eta_p^2 = 0.14$). Specifically, data from participants in the low algebra fluency group only revealed a statistically significant interference effect in the verbal WM condition ($M_{Control} = 1{,}257$ ms, $M_{Dual} =$

[2] Analyses involving the notation factor showed evidence of heterogeneity among the variances of the differences between possible pairs of levels of notation, and thus required a correction for sphericity. Thus, the Huynh-Feldt correction was used (although the correction did not change the statistical significance of any of the results).

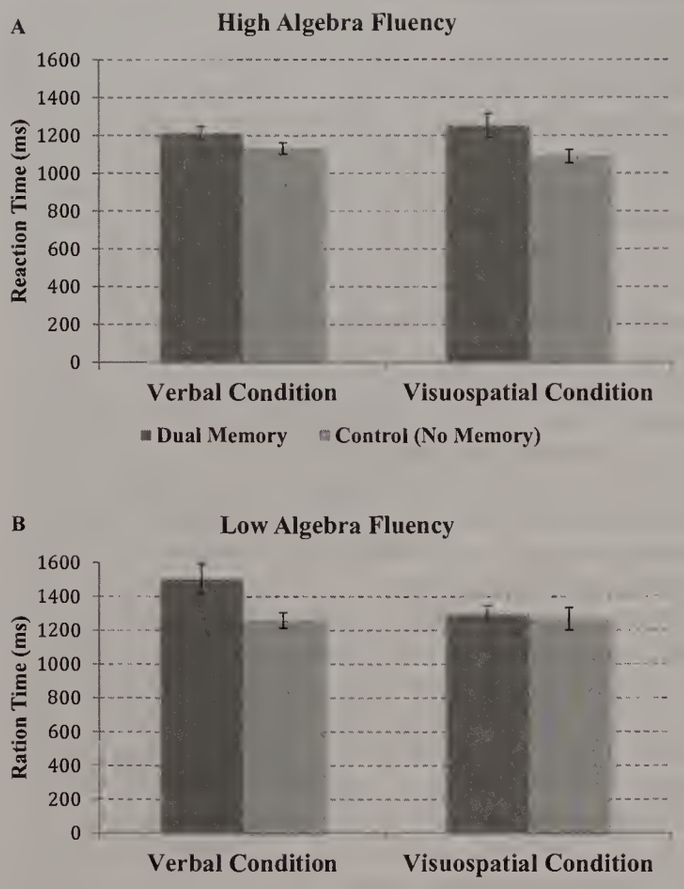


Figure 2. Reaction times on the magnitude comparison task across all notations (Fraction vs. Fraction [FvF], Decimal vs. Decimal [DvD], Decimal vs. Fraction [DvF], separated by working memory type (verbal vs. visuospatial) and interference (dual memory vs. no memory control). Individuals with high algebra fluency showed significant interference by both verbal and visuospatial working memory (WM; slower Reaction time [RT] in the dual relative to the control), whereas individuals with low algebra fluency showed interference only by verbal WM and not visuospatial WM.

1,501 ms; paired t test, $p < .001$) and not in the visuospatial condition (M_{Control} = 1,266ms, M_{Dual} = 1,293 ms; paired t test, $p = .6$).

Discussion

The current study used a dual-task paradigm to address adults' dominant strategy use during a rational number magnitude comparison task. We then explored the relationship between algebra ability and the differential engagement of distinct WM resources. Data revealed that, on average, adults tended to engage both verbal and visuospatial WM for both fraction and decimal comparison tasks. However, this pattern differed between individuals with higher and lower levels of algebra fluency.

Rational Number Notation

Although differences in the speed of processing distinct notation were found—such that decimal magnitudes were accessed significantly faster than those presented in fraction notation (as found in DeWolf et al., 2014; Hurst & Cordes, 2016a)—the use of visuospatial and verbal WM strategies did not vary as a function of notation type. Thus, we did not find evidence that the type of WM resources required to perform the task differed, on average, between fraction and decimal notation. Rather, adults seemed to use strategies that similarly relied upon verbal and visuospatial WM resources for both decimal and fraction notation and/or there was substantial individual variability in preferred strategy choice such that a dominant preference did not emerge. This is particularly striking given notable differences in structure between the two notations, claims that decimal notation is much more similar to whole number notation compared with fraction notation (e.g., Johnson, 1956), and findings that decimal magnitudes are more easily accessed than fraction magnitudes (the current study; DeWolf et al., 2014; Hurst & Cordes, 2016a). Despite these noted differences, results of our task did not reveal a distinction in the kinds of resources (visuospatial and verbal WM) recruited for fraction and decimal notation, suggesting that individuals may use similar types of strategies for both notations.

Although WM involvement differences across fractions and decimals were not obtained, the patterns of WM involvement when processing rational numbers found in the current study point to a potential distinction between whole numbers and non-whole-number rational numbers. Whereas both visuospatial and verbal WM were equally implicated when decimal and fraction magnitudes were judged, previous research suggests that magnitude judgments of whole numbers rely primarily on visuospatial resources suggesting that whole number magnitudes are spatially encoded (e.g., Simmons et al., 2011; van Dijck, Gevers, & Fias, 2009). Thus, despite parallels between decimal notation and whole number notation (Johnson, 1956), and despite evidence suggesting that all rational numbers (fractions, decimals, and whole numbers) are represented as falling along an integrated mental continuum in adults (Hurst & Cordes, 2016a), our data indicate that rational number magnitude processing may not perfectly parallel that of whole numbers. Instead, the involvement of verbal WM when processing decimal and fraction magnitudes indicates a role for symbolic calculation when adults process (nonwhole number) rational number magnitudes.

These findings shed light on cognitive models of rational number magnitude processing, suggesting that adults process rational number magnitudes more similarly to mental arithmetic (requiring both verbal and visuospatial WM resources) than to whole-number magnitudes (which likely require visuospatial resources and limited verbal resources). However, an interesting open question is how this pattern may change across development as a function of education. Research on WM recruitment during mental arithmetic tasks indicate a developmental trend such that children first rely primarily on visuospatial strategies, and then with greater experience, they use a mix of both visuospatial and verbal resources (e.g., McKenzie, Bull, & Gray, 2003; Raghubar et al., 2010). A similar developmental pattern may be found for fraction and decimal magnitude judgments as well. On the other hand, given how notoriously difficult rational number concepts are for children to acquire, coupled with the common whole number bias errors children show (e.g., treating fractions as two whole numbers as opposed to a coherent unit; Ni & Zhou, 2005), we may expect to see a reverse pattern of development in which children first learning rational numbers may initially show a greater reliance upon computational (verbal based) strategies when accessing rational

number magnitudes, with a later emerging reliance upon a mix of strategies. In fact, our finding that low algebra fluency adults relied primarily upon verbal strategies indicates that verbal strategies may be associated with low expertise in rational numbers and/or math ability more generally, consistent with the idea that children might initially have a primary reliance on verbal strategies.

Individual Differences in the Use of WM Resources

Results revealed that it was those adults with higher algebra fluency that were impacted by both verbal and visuospatial WM interference, but those with lower algebra fluency were only impacted by verbal WM interference. Although we did not directly measure the strategies employed by individual participants, assessing adults' reliance upon distinct WM resources allowed us to address the general kinds of strategies that may have been engaged during the rational number comparison task. In particular, the current findings suggest that those individuals who were relatively more fluent in algebra engaged both visuospatial-based strategies (i.e., reliance on the visuospatial sketchpad) and verbal strategies. However, lower algebra fluency was associated with the engagement of primarily verbal-based strategies (i.e., reliance on the phonological loop, but not the visuospatial sketchpad).

Given previous work looking at WM resource use during mental arithmetic, verbal WM interference in the current study is thought to be indicative of computational strategies requiring the memory of operands and/or interim solutions (see DeStefano & LeFevre, 2004). The use of verbal strategies, without also engaging visualization strategies (as was the case with low algebra individuals), may be particularly indicative of calculation or verbal rule-based strategies that do not also involve visualizing the magnitude or using complex calculations that involve mentally manipulating digits/components (which would involve both visuospatial and verbal WM). For example, these adults may have made comparisons based on component parts (i.e., numerators and denominators; tenths and thousandths place), converted fractions into decimals, or executed other verbal, calculation-based strategies. Importantly, however, our findings suggest that it is not simply the use of verbal WM during a rational number task that is associated with poor algebra fluency (because those with higher algebra fluency also engaged verbal resources), but rather a higher reliance on verbal WM with little reliance on visuospatial WM. This suggests that the engagement of particular kinds of strategies that rely primarily on verbal WM and not on visuospatial WM may be a critical predictor of poorer understanding of algebra. Those individuals, in the current study, who opted to use such computational strategies without also engaging visualization strategies likely have a poor understanding of how rational number symbols (in either decimal or fraction notation) translate to analog numerical magnitudes—a skill that may be important for success in algebra.

Individuals with high fluency, on the other hand, may have relied upon both visuospatial resources and verbal resources to process rational number magnitudes. The specific role of visuospatial WM is less clearly understood (relative to verbal WM) in the domain of mathematics. However, visuospatial WM has been implicated in visualization, such as in complex mental arithmetic that requires spatial movement (e.g., "carrying" in arithmetic; Raghubar et al., 2010). Thus, in the context of rational number processing, visuospatial and verbal WM engagement may be found when visualizing a proportional model (such as a pie chart or number line) or complex visual arithmetic (e.g., cross-multiplying, a strategy that presumably requires retaining verbal and spatial information).

Regardless, our data suggest that relying on both verbal and visuospatial strategies may be indicative of higher level conceptual processing. As such, this finding provides support for current policy recommendations that rational number instruction should highlight visuospatial representations of rational numbers as magnitudes (National Governors Association for Best Practices & Council of Chief of State School Officers, 2010; National Mathematics Advisory Panel, 2008). For example, one common recommendation is to emphasize the spatial organization of rational numbers along a number line, a recommendation drawn from other studies revealing training with number lines can be a successful intervention for understanding whole number magnitudes (Siegler & Ramani, 2009). Extending these recommendations, our results emphasize the importance of encouraging people to use strategies that incorporate both symbolic representations (requiring verbal resources) and visuospatial representations for thinking about rational number magnitudes. Although our findings cannot pinpoint exactly which strategies or representations adults in the high algebra fluency group engaged (e.g., pie charts, part–whole representations, number lines, discrete objects, complex visual arithmetic), they highlight the importance of engaging both visuospatial and verbal representations when thinking about rational numbers, rather than relying on exclusively verbal, calculation-based strategies and representations. As such, it may be that including many representations in the classroom may lead to a greater chance of the individual incorporating both visuospatial and verbal strategies. Future research should investigate which representations are most likely to be employed by mathematically fluent adults during rational number processing and, in addition, whether promoting particular visual representations in the classroom can lead to more efficient processing of rational numbers. Results of these studies will have implications both for understanding the cognitive processes underlying rational number processing, while also having important implications for educational practices.

Specificity to Algebra Fluency

Importantly, because rational number arithmetic fluency was entered as a covariate in our analysis, this pattern of WM resource use is not simply indicative of the speed of mathematical processing more generally. Rational number arithmetic fluency did predict the overall level of interference, suggesting that math fluency may be related to overall WM use. This is consistent with work suggesting that WM is involved in many areas of mathematics, including fraction conceptual knowledge and procedural ability (e.g., De Smedt, Verschaffel, & Conway et al., 2009; Geary, 2011; Jordan et al., 2013; Vukovic et al., 2014). Moreover, evidence suggests that as people gain practice with an activity, they require fewer WM resources to complete the activity (Gevins, Smith, McEvoy, & Yu, 1997; Jonides, 2004). Thus, those individuals who were most impacted in our dual-task (revealing the greatest amount of interference) were likely less practiced or fluent in rational number magnitude processing, suggesting that some aspect of our results may stem from overall differences in expertise. However, differences in the pattern of resource use (across verbal

and visuospatial WM) for those with differing levels of algebra fluency emerged even when controlling for performance on our rational number arithmetic assessment, indicating a specific relationship between rational number magnitude processing and algebra proficiency. That is, findings involving differences between individuals with relatively high and low algebra fluency do not reflect overall differences in cognitive or mathematical ability (which should be implicated about equally for Grade 8 algebra and rational number arithmetic), but rather are indicative of specific links between the processing of rational number magnitudes and algebra performance.

Limitations

There are some aspects of the current design that are worth noting. First, contrary to predictions, we did not find differences across rational number notation. One possibility is that the differences across notation are very small and that the current study did not have sufficient power to evaluate the three-way interaction required to see differences across notation within the current design. Alternatively, it may be that notation differences were not obtained due to our experimental design. The different notation trial types (FvF, DvD, and DvF) were intermixed within the same block, which may have impacted the types of strategies adults engaged in between notations. It is possible that if these distinct trial types were presented in separate blocks (i.e., all FvF trials were presented in a single block), then participants may have been more likely to settle upon a single strategy for working with the notation presented within that block of trials. If so, then notational differences may arise in contexts in which specific notations are consistent. Regardless, investigating differences across notation remains an open question for future research.

Additionally, although the current study used a variety of stimuli to investigate general processing of fraction and decimal magnitudes, it may be that the same strategies are not consistently used even within the same notation. For example, values on opposite sides of common bench marks, like 0.5 (or [1/2]) or 1, may lead to different strategies than comparing values on the same side of a common bench mark. In addition, there are several other component-based strategies (e.g., serially comparing decimals based on place value or comparing fractions based on numerators alone), heuristic-based strategies (e.g., choosing the longest decimal as the largest[3]), and format differences (e.g., vertical vs. horizontal alignment; presenting numbers one at a time instead of simultaneously) that may impact the kinds of strategies and resources that adults tend to engage. Because our study was not designed to specifically explore these issues, it was not possible to isolate the effects of these manipulations in our data, though this may be a topic for future research.

Lastly, our control task was designed in order to be perceptually and temporally identical to the dual task, and thus included verbal or visuospatial information between trials. Although adults were specifically instructed not to do so, it is possible that adults engaged some memory resources during the control task (i.e., there may be carry over effects between blocks). Importantly, however, our analyses do reveal significantly more interference during the dual tasks than the control tasks, making it unlikely that this greatly impacted our results.[4]

The Format of the Relationship

The current study leaves open the question of whether a *causal* relationship exists between the use of visuospatial strategies and algebra abilities. Given that children are taught rational number concepts prior to algebra, it may be that engaging complex visuospatial and verbal strategies when learning rational numbers may promote learning in more advanced math domains such as algebra. For example, representing rational numbers as holistic magnitudes (which would require symbolic and visuospatial strategies and not just verbal calculations) may be indicative of a more direct representation of the rational number system which algebra notation and manipulation is built (e.g., variables, unknown values, values along a line or curve). On the other hand, however, it may be that the learning of algebraic concepts can provide individuals with additional tools or strategies needed to engage both verbal and visuospatial strategies when processing rational numbers. For example, proficiency with algebraic manipulation (across both sides of an equation) may promote the use of visualization strategies when processing rational numbers as well. Lastly, it may be that visuospatial rational number strategies and high algebra ability are both associated with a third, general variable, such as visuospatial WM capacity or a general tendency toward abstract thinking. The correlational design of our study does not allow for a disentanglement of these accounts of mathematical learning. Therefore, future research should investigate this issue of causal direction in the relation between rational number processing strategies and algebra across various educational levels in order to better understand how these patterns may change across various stages in education and in young children whose WM capacities may not be at an adult level.

Conclusions

In sum, the current study used a dual-task WM paradigm to investigate individual differences in WM recruitment (verbal and visuospatial) during a rational number magnitude comparison task between individuals with relatively high and low algebra fluency. On average, adults were equally likely to engage visuospatial and verbal strategies when assessing the relative magnitudes of both decimals and fractions. Interestingly, however, individual variability in these strategies was associated with algebraic performance. Although individuals with relatively high algebra fluency relied on both verbal and visuospatial WM, individuals with relatively low algebra fluency relied more heavily on verbal WM to engage with rational numbers in both fraction and decimal notation. Thus, the use of strategies that involve both verbal and visuospatial resources (i.e., complex computations; visualization involving symbols) was associated with higher algebra performance, whereas using simple calculations or verbal rule-based strategies was as-

[3] In our study, 50% of trials were consistent with decimal length (meaning, longer decimal was the largest decimal), whereas 50% were inconsistent with length. In line with other work, there was an overall difference in performance, with inconsistent trials taking significantly longer than consistent trials ($p < 0.001$).

[4] Moreover, when we analyze each participant's first block only as a between-subject design, we get the same pattern of results involving WM interference and differences across algebra ability (although with much smaller sample sizes per cell).

sociated with lower algebra performance. These results add to the growing literature investigating the relationship between algebra ability and rational number understanding (e.g., Bailey et al., 2012; Hurst & Cordes, 2016b; Siegler et al., 2012), and further clarify the relationship by suggesting that individual differences in algebra ability may be associated with the use of different kinds of resource-based strategies in a rational number magnitude task. Additionally, results provide strong support for current recommendations to incorporate more visuospatial representations of rational number magnitudes, alongside symbolic representations, in the classroom (National Governors Association for Best Practices & Council of Chief of State School Officers, 2010; National Mathematics Advisory Panel, 2008), as the engagement of both verbal and visuospatial strategies was associated with advanced mathematical proficiency.

References

Baddeley, A. (1992). Working memory. *Science, 255,* 556–559. http://dx.doi.org/10.1126/science.1736359

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63,* 1–29. http://dx.doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation, 8,* 47–89. http://dx.doi.org/10.1016/S0079-7421(08)60452-1

Bailey, D. H., Hoard, M. K., Nugent, L., & Geary, D. C. (2012). Competence with fractions predicts gains in mathematics achievement. *Journal of Experimental Child Psychology, 113,* 447–455. http://dx.doi.org/10.1016/j.jecp.2012.06.004

Bonato, M., Fabbri, S., Umiltà, C., & Zorzi, M. (2007). The mental representation of numerical fractions: Real or integer? *Journal of Experimental Psychology: Human Perception and Performance, 33,* 1410–1419. http://dx.doi.org/10.1037/0096-1523.33.6.1410

Booth, J. L., & Davenport, J. L. (2013). The role of problem representation and feature knowledge in algebraic equation-solving. *The Journal of Mathematical Behavior, 32,* 415–423. http://dx.doi.org/10.1016/j.jmathb.2013.04.003

Booth, J. L., & Newton, K. J. (2012). Fractions: Could they really be the gatekeeper's doorman? *Contemporary Educational Psychology, 37,* 247–253. http://dx.doi.org/10.1016/j.cedpsych.2012.07.001

Booth, J. L., Newton, K. J., & Twiss-Garrity, L. K. (2014). The impact of fraction magnitude knowledge on algebra performance and learning. *Journal of Experimental Child Psychology, 118,* 110–118. http://dx.doi.org/10.1016/j.jecp.2013.09.001

Caviola, S., Mammarella, I. C., Cornoldi, C., & Lucangeli, D. (2012). The involvement of working memory in children's exact and approximate mental addition. *Journal of Experimental Child Psychology, 112,* 141–160. http://dx.doi.org/10.1016/j.jecp.2012.02.005

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12,* 769–786. http://dx.doi.org/10.3758/BF03196772

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122,* 371–396. http://dx.doi.org/10.1037/0096-3445.122.3.371

De Smedt, B., Verschaffel, L., & Ghesquière, P. (2009). The predictive value of numerical magnitude comparison for individual differences in mathematics achievement. *Journal of Experimental Child Psychology, 103,* 469–479. http://dx.doi.org/10.1016/j.jecp.2009.01.010

DeStefano, D., & LeFevre, J.-A. (2004). The role of working memory in mental arithmetic. *European Journal of Cognitive Psychology, 16,* 353–386. http://dx.doi.org/10.1080/09541440244000328

DeWolf, M., Grounds, M. A., Bassok, M., & Holyoak, K. J. (2014). Magnitude comparison with different types of rational numbers. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 71–82. http://dx.doi.org/10.1037/a0032916

Faulkenberry, T. J., & Pierce, B. H. (2011). Mental representations in fraction comparison. *Experimental Psychology, 58,* 480–489. http://dx.doi.org/10.1027/1618-3169/a000116

Geary, D. C. (2011). Cognitive predictors of achievement growth in mathematics: A 5-year longitudinal study. *Developmental Psychology, 47,* 1539–1552. http://dx.doi.org/10.1037/a0025510

Gevins, A., Smith, M. E., McEvoy, L., & Yu, D. (1997). High-resolution EEG mapping of cortical activation related to working memory: Effects of task difficulty, type of processing, and practice. *Cerebral Cortex, 7,* 374–385. http://dx.doi.org/10.1093/cercor/7.4.374

Hayes, J. R. (1972). On the function of visual imagery in elementary mathematics. In W. G. Chase (Ed.), *Visual information processing* (pp. 177–214). New York, NY: Academic Press.

Hitch, G. J. (1978). The role of short-term working memory in mental arithmetic. *Cognitive Psychology, 10,* 302–323. http://dx.doi.org/10.1016/0010-0285(78)90002-6

Holloway, I. D., & Ansari, D. (2009). Mapping numerical magnitudes onto symbols: The numerical distance effect and individual differences in children's mathematics achievement. *Journal of Experimental Child Psychology, 103,* 17–29. http://dx.doi.org/10.1016/j.jecp.2008.04.001

Hubber, P. J., Gilmore, C., & Cragg, L. (2014). The roles of the central executive and visuospatial storage in mental arithmetic: A comparison across strategies. *Quarterly Journal of Experimental Psychology (2006), 67,* 936–954. http://dx.doi.org/10.1080/17470218.2013.838590

Hurst, M., & Cordes, S. (2016a). Rational-number comparison across notation: Fractions, decimals, and whole numbers. *Journal of Experimental Psychology: Human Perception and Performance, 42,* 281–293. http://dx.doi.org/10.1037/xhp0000140

Hurst, M., & Cordes, S. (2016b). *The relationship between algebra and notation-specific rational number understanding in adults.* Manuscript submitted for publication.

Johnson, T. J. (1956). Decimal versus common fractions. *The Arithmetic Teacher, 3,* 201–203.

Jonides, J. (2004). How does practice makes perfect? *Nature Neuroscience, 7,* 10–11. http://dx.doi.org/10.1038/nn0104-10

Jordan, N. C., Hansen, N., Fuchs, L. S., Siegler, R. S., Gersten, R., & Micklos, D. (2013). Developmental predictors of fraction concepts and procedures. *Journal of Experimental Child Psychology, 116,* 45–58. http://dx.doi.org/10.1016/j.jecp.2013.02.001

Kilpatrick, J., & Izsak, A. (2008). A history of algebra in the school curriculum. In C. E. Greenes (Ed.), *Algebra and algebraic thinking in school mathematics* (pp. 3–18). Reston, VA: National Council of Teachers of Mathematics.

Koedinger, K. R., Alibali, M. W., & Nathan, M. J. (2008). Trade-offs between grounded and abstract representations: Evidence from algebra problem solving. *Cognitive Science, 32,* 366–397. http://dx.doi.org/10.1080/03640210701863933

Landy, D., Brookes, D., & Smout, R. (2014). Abstract numeric relations and the visual structure of algebra. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1404–1418. http://dx.doi.org/10.1037/a0036823

Maloney, E. A., Ansari, D., & Fugelsang, J. A. (2011). The effect of mathematics anxiety on the processing of numerical magnitude. *Quarterly Journal of Experimental Psychology (2006), 64,* 10–16. http://dx.doi.org/10.1080/17470218.2010.533278

McKenzie, B., Bull, R., & Gray, C. (2003). The effects of phonological and visuospatial interference on children's arithmetical performance. *Educational and Child Psychology, 20,* 93–108.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature, 215,* 1519–1520. http://dx.doi.org/10.1038/2151519a0

Moyer, R. S., & Landauer, T. K. (1973). Determinants of reaction time for digit inequality judgments. *Bulletin of the Psychonomic Society, 1,* 167–168. http://dx.doi.org/10.3758/BF03334328

National Governors Association Center for Best Practices & Council of Chief of State School Officers. (2010). *Common Core State Standards Initiative.* Retrieved from http://www.corestandards.com

National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* Washington, DC: U. S. Department of Education.

Ni, Y., & Zhou, Y. (2005). Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. *Educational Psychologist, 40,* 27–52. http://dx.doi.org/10.1207/s15326985ep4001_3

Raghubar, K. P., Barnes, M. A., & Hecht, S. A. (2010). Working memory and mathematics: A review of developmental, individual difference, and cognitive approaches. *Learning and Individual Differences, 20,* 110–122. http://dx.doi.org/10.1016/j.lindif.2009.10.005

Schneider, M., Grabner, R. H., & Paetsch, J. (2009). Mental number line, number line estimation, and mathematical achievement: Their interrelations in grades 5 and 6. *Journal of Educational Psychology, 101,* 359–372. http://dx.doi.org/10.1037/a0013840

Schneider, M., & Siegler, R. S. (2010). Representations of the magnitudes of fractions. *Journal of Experimental Psychology: Human Perception and Performance, 36,* 1227–1238. http://dx.doi.org/10.1037/a0018170

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., . . . Chen, M. (2012). Early predictors of high school mathematics achievement. *Psychological Science, 23,* 691–697. http://dx.doi.org/10.1177/0956797612440101

Siegler, R. S., & Ramani, G. B. (2009). Playing linear number board games—but not circular ones—improves low-income preschoolers' numerical understanding. *Journal of Educational Psychology, 101,* 545–560. http://dx.doi.org/10.1037/a0014239

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology, 62,* 273–296. http://dx.doi.org/10.1016/j.cogpsych.2011.03.001

Simmons, F. R., Willis, C., & Adams, A.-M. (2012). Different components of working memory have different relationships with different mathematical skills. *Journal of Experimental Child Psychology, 111,* 139–155. http://dx.doi.org/10.1016/j.jecp.2011.08.011

International Association for the Evaluation of Educational Achievement (IEA). (2005). *TIMSS 2003 Assessment.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education Boston College.

International Association for the Evaluation of Educational Achievement (IEA). (2013). *TIMSS 2011 Assessment.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education Boston College.

van Dijck, J.-P., Gevers, W., & Fias, W. (2009). Numbers are associated with different types of spatial information depending on the task. *Cognition, 113,* 248–253. http://dx.doi.org/10.1016/j.cognition.2009.08.005

Vukovic, R. K., Fuchs, L. S., Geary, D. C., Jordan, N. C., Gersten, R., & Siegler, R. S. (2014). Sources of individual differences in children's understanding of fractions. *Child Development, 85,* 1461–1476. http://dx.doi.org/10.1111/cdev.12218

Wu, H. (2001). How to prepare students for algebra. *American Educator, 25,* 10–17.

Appendix

Stimuli and Measures

Complete List of Magnitude Comparison Task Stimuli

	FvF	DvD	DvF
Verbal blocks	3/2 vs 5/7	.15 vs .168	3/2 vs .714
	4/3 vs 5/9	.196 vs .22	4/3 vs .56
	7/5 vs 2/3	.67 vs .594	7/5 vs .67
	7/4 vs 5/8	.835 vs .74	1.75 vs 5/8
	6/5 vs 4/9	.987 vs 1.12	1.201 vs 4/9
	3/5 vs 2/7	1.115 vs .99	3/5 vs .286
	2/5 vs 1/6	1.49 vs 1.687	.391 vs 1/6
	2/3 vs 1/4	1.54 vs 1.368	.67 vs 1/4
Visuospatial blocks	8/5 vs 2/3	.24 vs .256	8/5 vs .667
	9/5 vs 6/7	.33 vs .293	9/5 vs .86
	7/6 vs 5/9	.534 vs .47	1.19 vs 5/9
	6/5 vs 3/7	.639 vs .72	1.193 3/7
	8/5 vs 4/7	.98 vs 1.124	8/5 vs .57
	5/6 vs 3/8	1.075 vs 1.21	.83 vs 3/8
	4/9 vs 1/5	1.08 vs .948	.456 vs 1/5
	2/5 vs 1/7	1.493 vs 1.32	2/5 vs .143

Complete List of Rational Number Arithmetic Questions (8 Fraction and 8 Decimal)

Decimal Arithmetic Questions

$0.5 + 0.13 \quad 1.27 + 0.89 \quad 0.36 - 0.12 \quad 1.74 - 1.321$

$0.63 \div 0.12 \quad 1.452 \div 0.480.456 \times 0.32 \quad 1.75 \times 0.21$

Fraction Arithmetic Questions

$$\frac{3}{4} + \frac{7}{9} \quad \frac{2}{3} + \frac{5}{7} \quad \frac{4}{5} - \frac{4}{8} \quad \frac{6}{7} - \frac{1}{4}$$

$$\frac{8}{9} \times \frac{1}{3} \quad \frac{4}{5} \times \frac{3}{8} \quad \frac{4}{7} \div \frac{3}{5} \quad \frac{3}{9} \div \frac{3}{8}$$

Complete List of the 12 Algebra Questions

Question:

There are two pipes. The first pipe is x meters long. The second pipe is y times as long as the first one. How long is the second pipe?

Question:

In Zedland, total shipping charges to ship an item are given by the equation $y = 4x + 30$ where x is the weight in grams and y is the cost in zeds. If you have 150 zeds, how many grams can you ship?

Question:

Simplify the expression $2(x + y) - (2x - y)$

Question:

Give two points on the line $y = x + 2$

Question:

Simplify the expression $2a^2 \times 3a$

(*Appendix continues*)

Question:

The table below shows a relation between x and y

What is the relation between x and y?

Appendix (follows the sentence "The table below shows a relation between x and y" and before "What is the relation between x and y?").

x	1	2	3	4	5
y	1	3	5	7	9

Question: $3(2x - 1) + 2x = 21$ What is the value of x?

Question:

The number of jackets that Haley has is 3 more than the number Anna has. If n is the number of jackets Haley has, how many jackets does Anna have in terms of n?

Question:

$a = 3$ and $b = -1$ What is the value of $2a + 3(2 - b)$?

Question:

Joe knows that a pen costs 1 zed more than a pencil. His friend bought 2 pens and 3 pencils for 17 zeds. How many zeds will Joe need to buy 1 pen and 2 pencils?

Question:

Simplify the expression $4x - x + 7y - 2y$

Question:

If $\frac{x}{3} > 8$ then what does x equal?

Received January 12, 2016
Revision received November 6, 2016
Accepted November 10, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 5, 709–726

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000162

Phonological Processing in Children With Specific Reading Disorder Versus Typical Learners: Factor Structure and Measurement Invariance in a Transparent Orthography

Janin Brandenburg
German Institute for International Educational Research, Frankfurt am Main, Germany, and Center for Research on Individual Development and Adaptive Education of Children at Risk, Frankfurt am Main, Germany

Julia Klesczewski
Center for Research on Individual Development and Adaptive Education of Children at Risk, Frankfurt am Main, Germany, and Goethe University Frankfurt

Kirsten Schuchardt
University of Hildesheim

Anne Fischbach
German Institute for International Educational Research, Frankfurt am Main, Germany, and Center for Research on Individual Development and Adaptive Education of Children at Risk, Frankfurt am Main, Germany

Gerhard Büttner
Center for Research on Individual Development and Adaptive Education of Children at Risk, Frankfurt am Main, Germany, and Goethe University Frankfurt

Marcus Hasselhorn
German Institute for International Educational Research, Frankfurt am Main, Germany, Center for Research on Individual Development and Adaptive Education of Children at Risk, Frankfurt am Main, Germany, and Goethe University Frankfurt

Although children with specific reading disorder (RD) have often been compared to typically achieving children on various phonological processing tasks, to our knowledge no study so far has examined whether the structure of phonological processing applies to both groups of children alike. According to Wagner and Torgesen (1987), phonological processing consists of 3 distinct constructs: phonological awareness (PA), rapid automatized naming (RAN), and the phonological loop (PL) of working memory. The present study examined whether this phonological processing model which was originally developed for English orthography is also applicable to a more transparent language such as German. Furthermore, we tested whether the structure of phonological processing is invariant across typically achieving children and children with RD. Therefore, 209 German-speaking 3rd graders (100 typical learners and 109 children with RD) completed a comprehensive test battery assessing PA, RAN, and PL. Using confirmatory factor analyses, we compared the latent structure of these phonological processing skills across both groups. The study yielded 3 important findings: First, Wagner and Torgesen's (1987) model transfers to the German language and its orthography with transparent grapheme-to-phoneme correspondences. Second, the tripartite structure of phonological processing was evident across both groups (factorial invariance). Third, group invariance was also found for the measurement and structural components of the model (measurement invariance). These findings suggest that the nature

This article was published Online First November 28, 2016.

Janin Brandenburg, Department of Education and Human Development, German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany, and Center for Research on Individual Development and Adaptive Education of Children at Risk (IDeA), Frankfurt am Main, Germany; Julia Klesczewski, Center for Research on Individual Development and Adaptive Education of Children at Risk, and Institute of Psychology, Goethe University Frankfurt; Kirsten Schuchardt, Institute of Psychology, University of Hildesheim; Anne Fischbach, Department of Education and Human Development, German Institute for International Educational Research, and Center for Research on Individual Development and Adaptive Education of Children at Risk; Gerhard Büttner, Center for Research on Individual Development and Adaptive Education of Children at Risk, and Institute of Psychology, Goethe University Frankfurt; Marcus Hasselhorn, Department of Education and Human Development, German Institute for International Educational Research, Center for Research on Individual Development and Adaptive Education of Children at Risk, and Institute of Psychology, Goethe University Frankfurt.

This study was part of the longitudinal research project RAVEN (Differential Diagnostic Relevance of Working Memory in Children With Learning Disorders), exploring the developmental interplay between working memory and school achievement in children with learning disabilities. RAVEN is a multicentric study with data collection carried out in three federal states of Germany. We thank Christina Balke-Melcher and Claudia Mähler of the University of Hildesheim as well as Dietmar Grube and Claudia Schmidt of the University of Oldenburg for their involvement in carrying out this study. This research is part of the research initiative Developmental Disorders of Scholastic Skills, funded by the German Federal Ministry of Education and Research (Grant 01GJ1012A-D). The study was additionally funded by the IDeA Center, Frankfurt am Main.

Correspondence concerning this article should be addressed to Janin Brandenburg, Department of Education and Human Development, German Institute for International Educational Research, Schloßstraße 29, 60486 Frankfurt am Main, Germany. E-mail: brandenburg@dipf.de

of phonological processing is invariant across typically achieving children and children with RD acquiring the transparent orthography of German. Theoretical and practical implications are discussed.

Keywords: specific reading disorder, measurement invariance, phonological awareness, rapid automatized naming, phonological loop

It is well established that *phonological processing*—the ability to utilize the sound structure of oral language while processing written language (Wagner & Torgesen, 1987)—plays an essential role in the acquisition of reading and spelling skills. For instance, phonological processing contributes significantly to emergent literacy skills even when controlling for other crucial factors such as intelligence level (e.g., Babayiğit & Stainthorp, 2011; Schneider & Näslund, 1993), vocabulary and letter knowledge (e.g., Babayiğit & Stainthorp, 2011; Furnes & Samuelsson, 2011), or grammar skills (e.g., Nikolopoulos, Goulandris, Hulme, & Snowling, 2006). Moreover, longitudinal studies (e.g., Boscardin, Muthén, Francis, & Baker, 2008; Lambrecht Smith, Scott, Roberts, & Locke, 2008; Schneider & Näslund, 1993; Torppa et al., 2013; Wagner, Torgesen, & Rashotte, 1994) have demonstrated that phonological deficits are highly persistent over the childhood years and are causally related to the development of a reading disorder (RD). In fact, it is by now widely accepted that a *phonological core deficit* underlies the cognitive manifestation of an RD (Stanovich, 1988).

Compared to the vast amount of correlational and longitudinal studies conducted in the field, few have been dedicated to the structure of phonological processing. Quite recently, however, an increased interest can be found among researchers in uncovering the nature and dimensionality of phonological processing and examining the extent to which its conceptualizations are transferable to different (sub)populations. So far, this branch of research has mainly focused on the question to what extent the latent structure of phonological processing is transferable to different orthographies (e.g., Dutch: de Jong & van der Leij, 1999; Latvian: Sprugevica & Høien, 2004; Spanish: Anthony et al., 2006) and as to whether it applies to younger and older children alike (e.g., Anthony, Williams, McDonald, & Francis, 2007). Interestingly, despite the crucial role phonological processing plays in current models of RD no study has yet examined closely whether or not affected children show the same structure of phonological processing as their typically achieving peers. Therefore, the objective of this study was to determine and compare the factor structure and measurement invariance of phonological processing skills across these two groups of children.

On the Structure of Phonological Processing

Wagner and Torgesen (1987) developed an influential model of phonological processing which distinguishes three components: *Phonological awareness* (PA) describes a person's sensitivity to speech sounds and comprises the ability to analyze and manipulate the phonetic structure of spoken language. In this way, PA helps the beginning reader and speller to establish the crucial mapping between letters and sounds, necessary in acquiring an alphabetic script. *Phonetic recoding in working memory* (also referred to as the *phonological loop* [PL]) describes the temporary storage of verbal and acoustical information in working memory. For instance, during sentence reading the PL retains an acoustical representation of the words and thereby helps the reader to derive content and meaning. Likewise, the spelling process requires an acoustic representation of the to-be-written word in the PL. Finally, *phonological recoding in lexical access* (preferably assessed by a task called *rapid automatized naming* [RAN]) refers to how efficiently phonological information can be retrieved from the mental lexicon of long-term memory. More precisely, lexical access describes the mechanism by which a written word or another visual input leads to the rapid activation of its lexical entry through the process of phonological recoding (Wagner, 1986). Obviously, children with a large sight vocabulary who rapidly retrieve entire words are able to read and spell with greater efficiency than children who use an effortful letter-by-letter decoding strategy. Conceptualized in this way, RAN is considered to be a measure of the efficiency of visual-to-verbal recoding in the mental lexicon (cf. Wagner, 1986). Relatedly, Moll, Fussenegger, Willburger, and Landerl (2009) examined the RAN–literacy relationship in German and found evidence to suggest that RAN in this language may best be conceptualized as the automaticity of visual-verbal integration. Nevertheless, any conceptualization of RAN has to acknowledge that the naming of visual items involves a broad array of underlying processes: In addition to phonology, RAN also requires attention and inhibition as well as orthographic processing and general processing speed (cf. Kirby et al., 2010). Thus, although this study conceptualized RAN mainly as the phonological subcomponent that is responsible for visual–verbal integration, we additionally consider other theoretical explanations of the RAN–literacy relationship.

Wagner, Torgesen, and colleagues (Wagner, Torgesen, Laughon, Simmons, & Rashlotte, 1993; Wagner et al., 1994) further specified their model in later publications. For instance, they suggested that PA may actually consist of two discrete yet related factors, namely the ability to blend together phonemes or words (*phonological synthesis*) and the ability to identify and manipulate particular phonemes within words (*phonological analysis*). Although certainly this separation is of conceptual value, subsequent studies have supported neither this nor other multidimensional theories of PA (e.g., Anthony & Lonigan, 2004; Papadopoulos, Spanoudis, & Kendeou, 2009; Vloedgraven & Verhoeven, 2009). As a result, PA is nowadays generally considered to be a unitary ability. Moreover, Wagner et al. (1993, 1994) proposed to further subdivide lexical access into processes related to isolated versus serial naming. However, isolated naming has not been proven a unique predictor of literacy skills once serial naming is controlled (e.g., Logan, Schatschneider, & Wagner, 2011) and has thus been of only

minor importance in recent studies of reading. In line with these findings, the present study built on the originally proposed model that consists of three rather than five phonological processing components.

There is evidence that those phonological processing skills are related to each other in different ways: Statistical associations between PA and the PL are generally in the medium-to-high range, whereas their relations with RAN are often considerably smaller or even nonsignificant (see Norton & Wolf, 2012, for a review).

Given that Wagner and Torgesen's (1987) model was developed for English (an opaque orthography), one objective of the present study was to investigate whether the proposed structure also applies to the transparent orthography of German. In view of the differences between transparent and opaque orthographies, the question arose to what extent the findings for English can be generalized to other orthographies (e.g., Aro & Wimmer, 2003; Smythe et al., 2008; Ziegler, Perry, Ma-Wyatt, Ladner, & Schulte-Körne, 2003). By now, there is in fact some evidence from cross-language studies suggesting that the manifestation of RD is not universal but depends on the special characteristics of the underlying orthography (Landerl, Wimmer, & Frith, 1997; Vellutino, Fletcher, Snowling, & Scanlon, 2004).

Specific Reading Disorder

Specific RD (hereinafter just referred to as RD; ICD-10 code: F81.0) is a developmental learning disorder listed in the *International Classification of Diseases* (*ICD-10*) of the World Health Organization (WHO, 2011). The main feature according to definition is a significant and unexpected impairment in the development of reading and spelling skills: The learning problems are *significant* in that the child's performance is substantially below the level expected for the child's grade level; and they are *unexpected* because they contradict the child's intellectual potential. This uncoupling between intelligence and academic achievement (referred to as the *IQ discrepancy*) is at the heart of the medical definition (e.g., Ferrer, Shaywitz, Holahan, Marchione, & Shaywitz, 2010). Specifically, the IQ discrepancy has fueled the notion that RD is of neurobiological origin and arises from distinct and specific cognitive dysfunctions which are presumably but not exclusively located in the phonological domain (e.g., Lyon, Shaywitz, & Shaywitz, 2003). In other words, the medical definition of learning disorder argues that children who fulfill the IQ–achievement discrepancy criterion are qualitatively distinct from normal readers on the one hand, as well as from poor readers on the other hand, whose reading problems are in line with IQ expectations (e.g., Meyer, 2000).

Regarding the symptoms and manifestations of this disorder, ICD-10 acknowledges some differences between orthographies. While in English the problems tend to center on reading accuracy, there is by now ample evidence that in transparent languages such as German the main symptoms concentrate on slow and dysfluent reading (e.g., Landerl et al., 1997; Wimmer & Schurz, 2010). Moreover, because word recognition serves as a bottleneck for higher order reading skills, the children may also exhibit additional problems in text comprehension (e.g., Peterson & Pennington, 2015).

Phonological Processing in Children With Reading Disorder

Several meta-analytic studies (Kudo, Lussier, & Swanson, 2015; Melby-Lervåg, Lyster, & Hulme, 2012; Swanson, 2012; Swanson & Hsieh, 2009) and literature reviews (Mody, 2003; Vellutino et al., 2004) have come to the conclusion that children and adults with RD exhibit difficulties with all three components of phonological processing. Traditionally, most of the pertinent studies have been concerned with observed rather than latent variables and differences between groups are therefore mainly interpreted as functional deficits (cf. Schuchardt, Roick, Mähler, & Hasselhorn, 2008). That is, when children with RD perform lower on phonological tasks than their typically achieving peers, it is generally assumed that those performance differences are due to quantitative deficits in phonological processing. Yet, poorer performance on manifest measures may just as well result from structural differences in phonological processing: If the separation of phonological processing into three highly specialized components were not evident in children with RD, then PA, PL, and RAN would not contribute as much unique variance to the children's literacy development as they do in typical learners. From a theoretical point of view, this assumption of structural differences logically results from the medical definition, which conceptualizes RD as a distinct category rather than the lower end of the ability continuum.

However, this view has been challenged, raising the claim for a dimensional reconceptualization of RD (e.g., Branum-Martin, Fletcher, & Stuebing, 2013; Francis et al., 2005). Historically, Linda Siegel and Keith Stanovich were among the first who extensively criticized the IQ discrepancy criterion: Comparing IQ-discrepant with non-IQ-discrepant poor readers on a range of reading and cognitive measures, they found only little support for differences between the two groups, thereby questioning the usefulness of the medical definition (e.g., Siegel, 1989, 1992; Stanovich, 1994a, 1994b). Since then, there is an ongoing debate as to whether the cognitive differences between children with RD and typical learners are quantitative or qualitative in nature (e.g., Coghill & Sonuga-Barke, 2012; Compton, Fuchs, Fuchs, Lambert, & Hamlett, 2012). While the former view suggests that the cognitive characteristics of children with RD differ from normal reading only in level and degree, the traditional medical view assumes that the cognitive profiles also differ in pattern and kind.

To address this issue, a construct validation study that examines the equivalence of the underlying cognitive structure across the two groups seems reasonable, as this may provide empirical evidence in favor of either the dimensional or categorical conceptualization of RD. Moreover, given its crucial role for the development of reading and spelling, phonological processing is a cognitive source for which structural differences can be plausibly expected. Interestingly, to the best of our knowledge the assumption of structural equivalence of phonological processing has not yet been tested empirically. This constitutes a lack in current research because Schuchardt et al. (2008) and others (e.g., van de Schoot, Lugtig, & Hox, 2012) emphasized that results in a measurement instrument can only be validly compared across qualitatively distinct entities if the underlying latent constructs and the test properties do not differ systematically for these groups. If, however, group membership moderates either the nature of pho-

nological processing *or* the relationship that exists between the observed test scores and the latent constructs, valid interpretation of group differences is more difficult.

Conversely, if RD was quantitative in nature and just captured the lower end of the ability continuum, current practice would not be so problematic as—in this case—measurement invariance probably holds. Thus, instead of assuming that the structural conceptualization of phonological processing applies to children with RD in the same way as for typical learners, we argue that it is important to test for measurement and structural invariance among these two groups of children.

In fact, there are several circumstances that may suggest a difference in the latent structure of phonological processing in children with RD and typical learners: (a) Phonological processing may not follow a tripartite structure in children with RD; (b) the degree to which the phonological processing components vary may be different for children with RD and typical learners; (c) across the two groups, phonological processing tasks may not function equally well as indicators for the latent constructs; and (d) measurement error may systematically differ for children with RD and typical learners.

Phonological Processing May Not Follow a Tripartite Structure in Children With Reading Disorder

According to Carroll's (1993) hypothesis of proficiency-based divergence in latent factors, children might differ not only quantitatively in their actual cognitive performance levels, but also qualitatively in how their cognitive abilities are structured and specialized. Particularly, this hypothesis supposes that in young or low-proficient children cognitive abilities tend to be minimally differentiated and might therefore be captured by only one or few latent factors; however, with increasing age and proficiency the children's cognitive abilities fan out and become more and more specialized, which, in turn, produces a more complex factor structure. In line with this hypothesis, there is some evidence that children's phonological processing structure undergoes a developmental change during the first years of formal reading instruction. Specifically, the tripartite structure of phonological processing does not seem to be in place in children below the first and second grade. For instance, in a study conducted by Wagner et al. (1993), only lexical access (assessed by RAN tasks) emerged as a discrete ability in prereaders, whereas measures of PA and PL were not distinguishable from each other (i.e., they loaded on one and the same factor). Among the second graders tested in this study, it was however possible to fit the classical phonological processing model with separate factors for PA, PL, and RAN. This finding supports the notion that children's phonological abilities become increasingly differentiated when they develop from initial to skilled reading (cf. Lonigan et al., 2009). However, prereaders who later develop an RD might possibly not undergo this important developmental shift in their phonological processing structure—maybe due to some inherent neurobiological factors. This assumption seems reasonable as RD is considered a developmental disorder of neurobiological origin (e.g., Lyon et al., 2003; WHO, 2011). That is, PA and PL may constitute just one latent factor throughout the children's development and may be thus less differentiated, which may at least partially explain the children's learning problems.

The Degree to Which the Phonological Processing Components Vary May Be Different for Children With Reading Disorder and Typical Learners

The latent constructs of PA and RAN may be differently related to each other in children with RD and typical learners. For instance, in their meta-analysis of 49 correlational studies, Swanson, Trainin, Necoechea, and Hammill (2003) reported a moderate correlation of $r = .42$ between tasks of PA and RAN among skilled readers. In contrast, among children with poor reading skills the corresponding correlation—although corrected for range restrictions and sample size—was considerably lower ($r = .22$), indicating a lower linear dependence between PA and RAN in this group. These findings correspond with the *double-deficit hypothesis* (Bowers & Wolf, 1999). Accordingly, deficits in PA and in RAN constitute two independent sources of reading failure, resulting in different subtypes of struggling readers: Although the majority of poor readers (approximately 50% to 60%) is likely to exhibit deficits in both phonological processing skills (double deficit), others (approximately 25% to 35%) do poorly in either PA or in RAN only (single deficit); few poor readers may not be classified (Bowers & Wolf, 1999; Wolf et al., 2002). The existence of those subgroups may lead to lower relationships between latent conceptualizations of PA and RAN and may also lead to a higher variability of phonological processing skills in poor readers as opposed to good readers. Hence, tests of invariance seem especially suitable to compare the "true" (error-free) covariations and variance components of phonological processing skills across children with RD and typical learners.

Across the Two Groups, Phonological Processing Tasks May Not Function Equally Well as Indicators for the Latent Constructs

A particular phonological task may work as a good indicator of phonological processing in one group but not in the other. Statistically speaking, this would be the case if factor loadings varied as a function of group membership. For example, in their study with younger and older preschoolers, Lonigan et al. (2009) as well as Papadopoulos, Kendeou, and Spanoudis (2012) reported that the extent to which scores on PA tasks were accounted for by the underlying phonological factor differed systematically between the tested age groups. Additional evidence that phonological tasks may not be equally effective across different ability levels comes from Schatschneider, Francis, Forrman, Fletcher, and Mehta (1999). Using item response theory, the authors found that the difficulty and discriminability of various PA tasks was dependent on the children's ability levels. The authors therefore concluded that the diagnostic usefulness of a specific PA task cannot be a priori generalized across the ability continuum. It should thus be born in mind that one and the same phonological processing task may not function equally well across different (sub)populations. Some subtests may provide only limited information about low-ability children, but much information about high-ability children; whereas other subtests may be more effective in measuring phonological processing among children with low-ability levels. However, to the best of our knowledge no study has yet examined whether this restriction is also evident in children with RD. This is surprising, given that group invariance of factor loadings is an

important issue in construct validation: It ensures that the underlying latent construct is measured in the same way across groups and has thus the same underlying meaning (Meredith, 1993; Millsap, 2011).

Measurement Error May Systematically Differ for Children With Reading Disorder and Typical Learners

Lastly, the reliability of phonological tasks may systematically differ for both groups of children: Because struggling readers experience greater difficulties with phonological tasks than typical learners, they might get more easily distracted while performing those tasks or they might exhibit greater signs of frustration and motivation loss relative to their typically achieving peers. Likewise, floor effects may also occur more often in children with RD. Obviously, those effects might reduce the reliability of the testing instruments. To rule out the possibility that measurement error systematically differs across groups, it seems therefore necessary to examine whether residual variances of phonological processing tasks (i.e., the observed variance *not* accounted for by the underlying phonological processing factors) are invariant across children with RD and typical learners.

To summarize, rather than assuming that the latent structure of phonological processing applies to children with RD and typical learners alike, we argue that it is important to test for structural and measurement invariance among these two groups of children. Particularly, our study was designed as a construct validation study that aimed at answering the following two questions:

> *Research Question 1:* Does phonological processing follow the proposed tripartite structure in children with RD who acquire a transparent orthography?
>
> *Research Question 2:* If so, is the phonological processing model invariant across children with RD and typical learners?

Method

Procedure

The study was part of a multicentric research project that aimed at investigating the interplay between cognitive functioning and school achievement in children with learning disorders.

Recruitment of participants. All children of the reference group and most children of the RD group (92%) were recruited via a screening on school achievement that took place in elementary schools in and around three cities located in the northern and central parts of Germany (viz., Frankfurt am Main, Hildesheim, and Oldenburg). The remaining children with RD were recruited by a counseling center for learning disabilities in Hildesheim. Overall, 3,205 children (age: M = 8 year 7 months, SD = 6 months; 49.2% girls) from 134 schools and 280 classes participated in the screening. The children completed standardized school achievement tests of reading, spelling, and mathematics as well as a standardized measure of nonverbal IQ. They were tested in groups at their school over two 90-min lessons. Given our diagnostic criteria (outlined below), the screening revealed a prevalence rate of RD of about 8.7% (Fischbach et al., 2013), which is within the range reported in other prevalence studies (e.g., Dirks, Spyer, van Lieshout, & de Sonneville, 2008; Landerl & Moll, 2010; see also Peterson & Pennington, 2015, for a review). Children fulfilling the criteria for the RD group or the reference group were either invited to take part in the main study or a related study. Of the 527 children invited to the reference group, about 30% decided to participate in one of the two studies. Overall, participating children were mostly comparable to nonparticipating children with respect to the classification measures and age. Significant but marginal group differences only emerged in reading (T score difference: ΔM = 1.50; ΔSD = 0.03; η_p^2 = .018) and in mathematics (T score: ΔM = 1.61; ΔSD = −0.26; η_p^2 = .017), with nonparticipating children outperforming participating children. For the RD group, return rate was about 34% and again participating children were mostly comparable to nonparticipating children. Significant but marginal group differences emerged in reading (T score: ΔM = 3.05; ΔSD = 0.43; η_p^2 = .029) and in spelling (T score: ΔM = −1.18; ΔSD = −0.91; η_p^2 = .011), with nonparticipating children performing slightly better in reading but poorer in spelling.

Assessment of cognitive functioning (main study). Measures of phonological processing were only administered to students participating in the main study. The assessment took place individually in two sessions, each lasting up to 90 min. Student research assistants tested the children in schools or in the universities' laboratories. Parental informed written consent was obtained for all children prior to testing. Participation was voluntary and consent could be withdrawn at any time without giving reasons.

Participants

The sample of the main study included 209 third graders. Of these, 100 children belonged to the reference group and 109 to the RD group. Classification was based on norm-referenced measures (standard scores). The criteria for the reference group were as follows: Children's achievement scores in reading, spelling, and mathematics were at grade expected levels, with standardized scores of at least $T \geq 45$ (T score: M = 50; SD = 10). In contrast, children of the RD group showed achievement scores in reading and/or spelling that were below grade expected levels (i.e., at least 1.0 SD below the normed reference group's mean; T score ≤ 40), whereas their performance in mathematics was grade appropriate ($T \geq 40$ and at least 5 T points above the child's reading and spelling scores). In line with ICD-10 (WHO, 2011), we additionally applied an IQ–achievement discrepancy criterion to the definition of RD: In particular, the children showed a critical discrepancy of at least 1.2 SDs between their nonverbal IQ and their literacy achievement. Also, all children showed at least average intelligence (IQ $\geq$ 85).

Because the cut-off criteria used in the classification of RD are rather heterogeneous (e.g., Elliott & Grigorenko, 2014), we want to outline the rationale for the criteria used in the present study: In general, a norm-referenced cut-off score of $T < 40$ for the low achievement criterion and of 1.2 SDs for the IQ discrepancy criterion correspond to the recommended diagnostic guidelines (Strehlow & Haffner, 2002) and they are most frequently used in German educational and clinical settings (Hasselhorn, Mähler, & Grube, 2008; Klicpera, Schabmann, & Gasteiger-Klicpera, 2010).

That is, by applying those cut-off scores, our sample best represented the subpopulation of schoolchildren in Germany commonly referred to as having an RD.

Table 1 shows the descriptive characteristics of the sample as a function of group. A multivariate analysis of variance was performed to check whether the groups differed with respect to the classification measures. For these and the following statistical analyses, the alpha level was set at $p = .05$. Effect sizes are reported using partial eta-squared (η_p^2) classified by Cohen (1988) as small (.01–.05), medium (.06–.13), and large (≥.14) effects. The multivariate main effect was statistically significant, Wilks's $\lambda = .29$, $F(4, 204) = 125.33$, $p < .001$, $\eta_p^2 = .71$; and was therefore followed up with separate univariate analyses of variance: No statistically significant differences between groups were found for nonverbal intelligence, $F(1, 207) < 1$, $MSE = 108.69$, $p = .402$, $\eta_p^2 < .01$; and for mathematical skills, $F(1, 207) < 1$, $MSE = 36.65$, $p = .493$, $\eta_p^2 < .01$. However, as could be expected due to the sampling procedure, groups differed significantly in reading skills, $F(1, 207) = 193.54$, $MSE = 50.34$, $p < .001$, $\eta_p^2 = .48$; and in spelling skills, $F(1, 207) = 263.07$, $MSE = 36.61$, $p < .001$, $\eta_p^2 = .56$. There were no statistically significant group differences with respect to chronological age as indicated by an analysis of variance, $F(1, 207) = 2.59$, $MSE = 28.84$, $p = .109$, $\eta_p^2 = .01$. Whereas sex distribution was balanced in the reference group (50 boys; 50 girls), there were more boys than girls in the RD group (77 boys; 32 girls), $\chi^2(1, N = 209) = 9.32$, $p = .002$. This is in line with prevalence studies showing that learning disorders in the literacy domain are generally more frequent in boys than in girls (e.g., Moll, Kunze, Neuhoff, Bruder, & Schulte-Körne, 2014). Lastly, the RD group and the reference group were balanced with respect to the children's home towns, $\chi^2(2, N = 209) < 1$, $p = .952$ (reference group: 46% Frankfurt, 29% Hildesheim, 25% Oldenburg; RD group: 44% Frankfurt, 29% Hildesheim, 27% Oldenburg).

Reading and Writing Curriculum in German Elementary Schools

Due to the transparency of the German orthography, reading and spelling instruction often follows a synthetic phonics-based teaching approach. That is, children are systematically taught all existing grapheme–phoneme relations and learn how to derive word pronunciation on the basis of these conversion rules. Likewise, there is a high prominence on phoneme-grapheme relations in writing instruction, at least in the first years of schooling: Children are usually taught to segment the sound sequences of words into individual phonemes and are encouraged to write down the corresponding graphemes (e.g., Niedersächsisches Kultusministerium, 2006).

Tasks and Materials

Classification measures. To obtain an estimate of general cognitive ability, children completed the German version of the *Culture Fair Intelligence Test 1* (CFT 1; Cattell, Weiß, & Osterland, 1997). The CFT 1 is a nonverbal measure used as an

Table 1
Descriptive Statistics for All Measures as a Function of Group

Measures	Reference group (n = 100)				RD group (n = 109)			
	M	*SD*	Skew.	Kurt.	*M*	*SD*	Skew.	Kurt.
Classification measures (independent variables) and age								
Age (in months)	102.16	4.86	−.43	−.21	103.36	5.80	.60	.02
Intelligence[a]	106.86	11.19	.77	.52	108.07	9.68	.96	1.17
Mathematics[b]	53.98	5.49	.42	−.43	53.40	6.53	.06	−.83
Reading[b]	53.53	5.91	.95	.99	39.86	8.03	.75	−.46
Spelling[b]	51.36	5.75	.92	.27	37.77	6.31	.58	−.01
Phonological processing measures (dependent variables)								
Phonological awareness								
Vowel length	4.63[c]	2.76	−.05	−.95	3.45[c]	2.25	.31	−.58
Vowel substitution	9.46[d]	2.28	−1.04	.76	7.06[c]	2.94	−.51	−.37
Phoneme reversal	9.41[c]	4.79	−.04	−1.25	5.21[d]	4.16	.84	.05
Phonological loop								
One-syllable word span	3.95	.65	.08	−.02	3.74	.58	−.27	−.52
Three-syllable word span	3.12	.44	.37	.45	3.00[d]	.37	.17	.30
Digit span	4.61	.60	−.41	.70	4.17	.58	−.36	−.14
Serial naming (in s)								
Color naming	49.93	10.20	.78	.81	51.86[e]	10.56	.82	.33
Digit naming	28.88[c]	6.20	.94	1.04	34.25[f]	7.29	.62	−.06
Letter naming	31.30	6.87	1.16	1.34	34.35[g]	7.18	1.10	1.72
Object naming	47.42[d]	7.53	.54	.23	48.29[f]	7.41	.63	.96

Note. RD = reading disorder; Skew. = skewness; Kurt. = kurtosis; Phon. = phonological. Overall, less than 1.5% of the phonological processing data were missing. There were various reasons for missing data, which can best be summarized as technical problems or test administration errors that occurred during the testing (e.g., children did not understand the test correctly or did not complete the task, technical problems with data recording especially during the RAN test, student research assistants did not administer the test correctly).
[a] IQ score. [b] *T* score. [c] Data of one participant are missing. [d] Data of two participants are missing. [e] Data of three participants are missing. [f] Data of five participants are missing. [g] Data of four participants are missing.

indicator of fluid intelligence; it comprises five subtests that can be classified according to two parts. The first part consists of two measures, which assess perceptual speed, visual attention and visuomotor ability. The second part consists of three subtests, which examine deductive reasoning skills and figural thinking. All items are dichotomously scored and then combined to an overall performance score. Split–half reliability is .90 and .92 for the age groups studied. The validity of the CFT 1 has been well established and the manual states reasonable levels of convergent and criterion validity as well as factorial validity.

We tested the children's reading skills with Ein Leseverständnistest für Erst- bis Sechstklässler (ELFE 1-6; Lenhard & Schneider, 2006), which consists of three subtests. The first subtest assesses decoding speed by means of a picture–word matching procedure: Each of the 72 items consists of one picture and four written words. The children's task is to identify the word that corresponds to the picture. The children have 3 min to work on as many items as possible. The second subtest measures reading at sentence level and consists of 28 unrelated sentences, each missing a word. Out of a set of five words, the children are asked to identify the word that completes each sentence correctly. Children have 3 min to work on as many items as possible. The third subtest assesses reading at text level and requires the children to answer multiple-choice questions in response to short narratives. Children have 7 min to complete the task. The ELFE 1–6 is designed as a speed test rather than a power test so that item difficulty (especially for Subtests 1 and 2) is generally low. Consequently, children rarely make mistakes and skill differentiation is mainly based on reading speed. A reading test covering reading speed and reading comprehension was used rather than a reading accuracy test, as reading accuracy is usually high in transparent orthographies (e.g., Landerl, 2001; Landerl et al., 1997), and it consequently does not distinguish sufficiently between good and poor readers. Items of the three subtests are dichotomously scored and then combined to an overall performance score. Internal consistency of the ELFE 1–6 is high with values between .92 and .97. The manual reports adequate levels of convergent validity (e.g., correlations of $r = .48$ to .79 with other standardized reading tests) and satisfactory levels of criterion validity (e.g., correlations of about $r = -.76$ with the children's grades in German).

To assess spelling achievement, the Weingartener Grundwortschatz Rechtschreib-Test für zweite und dritte Klassen (WRT 2+; Birkel, 2007), a German spelling test for second and third graders was administered. This test requires children to spell 43 dictated words embedded in short sentences. Items of the WRT 2+ are dichotomously scored; the dependent variable is the number of correct spellings. Internal consistency for this measure is reported as high with Cronbach's $\alpha = .94$. The manual reports satisfactory levels of convergent validity (e.g., correlations of $r = .65$ to .85 with other standardized spelling tests) and adequate levels of criterion validity (e.g., correlations of $r = -.65$ to $-.69$ with the children's grades in German).

To control for co-occurring learning disabilities in mathematics, children completed the Deutscher Mathematiktest für zweite Klassen (DEMAT 2+; Krajewski, Liehm, & Schneider, 2004), a curricular-valid test of basic arithmetic, magnitude, and geometry. Items of this test are dichotomously scored; the dependent variable is the number of problems solved correctly. Internal consistency of the DEMAT 2+ is reported as .91 for third graders. The DEMAT 2+ shows moderate to strong levels of prognostic validity ($r = .63$ to .67), convergent validity ($r = .53$ to .67) and criterion validity ($r = .66$).

Measures of phonological processing. Because this study can be construed as a construct validation study, we particularly used those types of tasks that are commonly considered to be typical measures of phonological processing. Following the suggestions of Anthony et al. (2007) as well as Lonigan et al. (2009), we further deemed it necessary to only use those measures that are not related with two or more phonological processing skills at the same time to prevent confounding of constructs. For instance, a nonword repetition task was therefore not applied in the present study. Although nonword repetition was originally introduced as a relatively pure measure of the PL (Gathercole & Baddeley, 1993), researchers have since argued that the mechanisms underlying repetition of nonword tasks are by far more complex (e.g., Archibald & Gathercole, 2007; Gathercole, 2006) and less well understood (cf. Metsala, 1999) than in serial recall measures such as word span or digit span. Furthermore, it is by now widely accepted that (other than serial recall) nonword repetition (a) assesses the phonological *sound quality* of item storage rather than the overall *storage capacity* of the PL (e.g., Hasselhorn, Grube, & Mähler, 2000) and (b) it is thus in a conceptual sense highly related to PA (see Bowey, 1997; Gathercole, 2006, for reviews). Specifically, storing and repeating a nonword for which no lexical entry exists requires children to identify and segment the presented phoneme structure in a deep manner, because only in this way the item can be maintained correctly in the PL. In fact, empirical studies (e.g., Metsala, 1999) have demonstrated that nonword repetition is not only accounted for by phonological storage capacity, but also by phonological sensitivity as assessed in PA.

Along with the description of subtests, we provide sample-based reliability estimates. Overall, the internal consistency of the measures was around .70 to .90, which is within an acceptable and common range for basic research (cf. Nunnally & Bernstein, 1994).

Phonological awareness (PA). The Basiskompetenzen für Lese-Rechtschreibleistungen (BAKO 1–4; Stock, Marx, & Schneider, 2003) was used to assess the children's PA on phoneme level. Of the seven subtests of the BAKO 1–4, we selected those three subtests that showed the best item characteristics, namely Vowel Substitution, Vowel Length, and Phoneme Reversal. That is, not used in the present study were the BAKO 1–4 subtests Nonword Segmentation, Phoneme Inversion, Sound Categorization, and Phoneme Deletion. The manual reports expected and satisfactory levels of criterion validity by correlating the BAKO 1–4 with a range of widely used achievement tests (e.g., moderate to high correlations with measures of reading and spelling; nonsignificant to low correlations with nonverbal IQ). Items of the BAKO 1–4 are dichotomously scored.

In the Vowel Substitution test, the child's task is to substitute all /a/ vowels in a given word by an /i/ vowel (e.g., Sand → Sind). This test is based on vowel phonemes rather than consonant phonemes, because in a lot of languages (including German) vowel changes are used to indicate the tenses of irregular verbs and are thus a crucial phonological marker (cf. Wimmer, Landerl, Linortner, & Hummer, 1991). The test consists of 12 items (eight words, four pseudowords), which range between two and four syllables in length. Three practice trials are presented prior to testing. Internal

consistency of this measure (based on the Kuder–Richardson formula 20 (KR_{20}) due to the dichotomy of the item scoring) was KR_{20} = .78 for the RD group and KR_{20} = .76 for the reference group.

The Vowel Length subtest, which has mainly been used in transparent orthographies, is a modification of Bradley and Byrant's (1985) sound categorization/oddity detection task and assesses vowel *duration* instead of vowel *identity*. Accordingly, this test is phonologically more demanding than its original and it thus works well with third graders (Stock et al., 2003). In contrast, discrimination of vowel *identity* (as assessed in the conventional sound categorization task) is mastered early by children acquiring transparent orthographies such as German (Landerl, 2003). Moreover, a measure of vowel length is of interest because being able to correctly perceive and discriminate vowel lengths in spoken German is an important phonological skill, which is, for example, required in learning the difficult German spelling rules to mark short and long vowels (cf. Landerl, 2003). The child is presented four pseudowords, all of which contain the same vowel phoneme in the middle. Yet, one of the pseudowords differs from the others in vowel length (e.g.,/re:m/-/fe:r/-/nɛl/-/be:f/). The child's task is to identify the item that does not match the others. The subtest consists of two practice trials and 10 test trials. Internal consistency of this measure was KR_{20} = .69 for the RD group and KR_{20} = .79 for the reference group.

In the Phoneme Reversal subtest, the child's task is to pronounce a given (pseudo)word in reversed order (e.g., ruf → fur). Of the 18 test items, 10 consist of pseudowords, four of which turn to real words after reversal. Again, two practice trials are presented prior to testing. Internal consistency of this measure was KR_{20} = .88 for the RD group and KR_{20} = .89 for the reference group.

Items of the BAKO 1–4 were presented audibly via computer and the examiner recorded the child's verbal responses on a protocol sheet. Subtest presentation was stopped once the child answered three subsequent items incorrectly. The dependent variable was the number of correct trials up to the point at which the subtest was stopped.

Phonological loop (PL). Three subtests of the computerized and adaptive Arbeitsgedächtnistestbatterie für Kinder von 5 bis 12 Jahren (AGTB 5–12; Hasselhorn et al., 2012) were administered to assess phonetic recoding in short-term memory. The AGTB 5–12 is a German computerized test battery, which assesses working memory skills according to Baddeley's (1986) multicomponent model. Construct validity of the AGTB 5–12 was established in a large study with 1,669 children (Michalczyk, Malstädt, Worgt, Könen, & Hasselhorn, 2013). Furthermore, the AGTB 5–12 demonstrates significant criterion validity with respect to reading and spelling tests (Hasselhorn et al., 2012).

In the digit span task, increasing sequences of different digits are presented audibly at the rate of one digit every 1.5 s. No digit appears twice in a particular sequence. The child's task is to repeat the sequence orally in the same serial order as presented. The sample Cronbach's alpha coefficient was .90 for the RD group and .92 for the reference group.

Similarly, the word span task requires the serial repetition of high-frequency words, which are presented audibly at the rate of one word every 1.5 s. Word sequences are constructed out of nine phonologically and semantically dissimilar German nouns. Each word appears only once within a particular sequence. There are two versions of the task—one using monosyllabic and one using trisyllabic words—resulting in separate span scores for short and long words, respectively. For the monosyllabic word span, the sample Cronbach's alpha coefficient was .89 for the RD group and .92 for the reference group. For the trisyllabic word span, Cronbach's alpha was .79 and .89, respectively.

All three tasks are span measures with an adaptive testing procedure. They consist of 10 trials, divided into five testing blocks with two trials each. The first testing block starts with a three-item sequence, and sequence length is adjusted after each response: If the child recalls the presented trial correctly, the sequence length of the subsequent trial increases by one item. If, however, the child's recall is incorrect, the sequence length of the next trial decreases by one item. In the remaining four testing blocks, sequence length is adjusted more conservatively as follows: If the child recalls both trials of the testing block correctly, the span length of the next block increases by one item. If, however, the child recalls both trials incorrectly, the span length decreases by one item. If recall is incorrect for only one of the two trials, the span length remains the same. The calculation of the span score is based on the mean performance in the last four testing blocks: For each correct response, the child receives a score that corresponds to the span length. For instance, if the child correctly recalls a five-item sequence, he or she receives five points. A false response is assigned the span length decreased by one item (e.g., incorrect repetition of a five-item sequence results in four points only).

Rapid automatized naming (RAN). This task measured lexical access from long-term memory. The RAN task consisted of two alphanumeric subtests, which assessed naming speed for digits and letters, with items adapted from the Differentiaal Diagnostiek van Dyslexie (3DM Dyslexie; Blomert & Vaessen, 2008). In addition, two nonalphanumeric subtests assessed naming speed for colors (red, yellow, blue, green, and black; items adapted from Denckla & Rudel, 1976), and objects (car, fish, hammer, dog, candle; items adapted from Blomert & Vaessen, 2008). In each subtest, items are arranged randomly in five rows of 10 on a white paper (size: 41.0 × 29.5 cm). The child's task is to name all items as quickly as possible while making as few errors as possible. Naming time (in seconds) served as the dependent variable, that is, lower scores indicated higher performance. Each subtest was preceded by a short practice trial (i.e., two rows à five items) to familiarize the child with the material. The sample Cronbach's alpha coefficient was .73 for the RD group and .74 for the reference group.

Statistical Analyses

Research Question 1: Factor structure of phonological processing. To investigate the factor structure of phonological processing in children with RD versus typical learners, two alternative models were tested: (a) a two-factor oblique model in which the RAN items were captured by one of the factors, whereas the PA and PL items were captured by the other factor, and (b) a three-factor oblique model with separate latent factors for PA, PL, and RAN.

The two-factor model is nested under the three-factor model, because it can be obtained by restricting the intercorrelations between the latent factors of PA and PL to 1. Models were

tested in Mplus Version 7.11 (Muthén & Muthén, 1998-2012) using maximum likelihood estimation with robust standard errors (MLR). The MLR estimator treats missing values in a full information approach and can be applied to non-normal distributed data (Wang & Wang, 2012). Following Hu and Bentler's (1999) criteria of model fit, a good fit to the data was indicated by (a) a comparative fit index (CFI) with values of at least .95, (b) a root mean square error of approximation (RMSEA) with values of .06 or less, and (c) a nonsignificant chi-square test. In addition, we report two information criteria indices, namely the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), where smaller values represent a better fitting model.

Research Question 2: Measurement invariance of phonological processing. In construct validation studies, the following invariance tests are generally considered necessary: invariance of factor loadings, variances, and covariances (e.g., Byrne, 2012; Vandenberg & Lance, 2000). Thus, the next set of analyses determined whether the measurement parameters (i.e., the factor loadings and residual variances) as well as the structural parameters (i.e., the factor variances and covariances) of the established factor model operated equivalently across both groups of children. Hence, a sequence of increasingly restrictive multigroup models was tested, with equality constraints imposed to the covariance structure in a hierarchical manner. In each testing step, we used nested model comparison to assess whether the more restrictive model was preferable to the less restrictive but slightly better fitting comparison model. Specifically, nested model comparison was twofold: (a) Because we used the MLR estimator, changes in model fit were examined with the Satorra–Bentler scaled chi-square difference test (ΔSB-χ^2; Satorra & Bentler, 2001) rather than the likelihood ratio test. A nonsignificant value of the ΔSB-χ^2 statistic implies that the restrictive model fits the data just as well as the less restrictive comparison model (e.g., Wang & Wang, 2012). Thus, the hypothesis of parameter invariance can be accepted. (b) The CFI difference value (ΔCFI) served as a second indicator for multigroup invariance: A difference value less than or equal to 0.01 between the restricted and the comparison model suggests parameter invariance (Cheung & Rensvold, 2002).

Results

Data Screening

Table 1 shows descriptive statistics of the phonological processing measures as a function of group. Prior to the main analyses, we evaluated whether the data met basic assumptions of confirmatory factor analyses. In particular, none of the zero-order correlations between the manifest variables was above .80 (see Table 2), indicating no problem with multicollinearity. In addition, the data were checked for univariate outliers that we classified in terms of cases more than 3.5 *SD*s from the sample's means: Of the 2,067 values in the dataset, seven values were univariate outliers (four children of the RD group; three children of the reference group). These values were deleted from further analyses. No cases were identified as multivariate outliers with $p < .001$ through Mahalanobis distance. There was no evidence that the assumption of univariate normality distribution was violated, because all measures showed skewness less than 3 and kurtosis less than 4. Yet, Mardia's (1974) test of multivariate normality revealed a violation of the multivariate skewness assumption, whereas assumption of multivariate kurtosis was met for both groups. Model estimation was therefore based on the MLR estimator, which offers chi-square test statistics and standard errors that are robust to nonnormal data (Wang & Wang, 2012).

Research Question 1: Measurement Model of Phonological Processing

For both groups, the two-factor oblique model provided only poor fit to the data, as can be seen in Table 3. We then tested the three-factor oblique model: Whereas this model provided an excellent fit to the data of the RD group, results of the reference group revealed a poor fitting model. Note, for example, that the chi-square value was highly significant in the reference group. In relatively small samples such as ours a significant chi-square value is always of concern (Kline, 2016). We therefore consulted modification indices, which yielded evidence that the residuals between the object naming task and the color naming task covaried highly in the reference group (modification index > 22). This covariation may be due to method effects, as both measures are

Table 2
Correlations Among Phonological Processing Measures

Measures	1	2	3	4	5	6	7	8	9	10
1. Vowel substitution		.14	.30	.21	.22	.14	.11	.16	.16	.18
2. Vowel length	.19		.39	.25	.26	.27	−.16	−.10	−.08	.00
3. Phoneme reversal	.40	.22		.06	.15	.09	−.06	.06	.03	.04
4. One-syllabic word span	.21	.08	.10		.63	.65	.05	−.13	−.10	−.08
5. Three-syllabic word span	.20	.22	.28	.47		.67	.07	−.02	.03	.01
6. Digit span	.20	.17	.17	.57	.47		−.04	−.13	−.05	−.03
7. Object naming	−.13	−.10	−.03	−.21	−.09	−.07		.60	.29	.35
8. Color naming	−.09	−.01	−.06	−.06	−.04	−.08	.50		.36	.42
9. Digit naming	.04	.01	−.08	−.12	.07	−.08	.39	.47		.57
10. Letter naming	−.07	−.16	−.13	−.02	.10	−.04	.25	.29	.44	

Note. Intercorrelations for the reference group are presented above and those for the RD group are presented below the diagonal. All correlation coefficients of $r \geq .20$ are significant at $p < .05$.

Table 3
Measurement Model of Phonological Processing as a Function of Group

	Reference group							RD group						
CFA models	χ^2	*df*	*p*	CFI	RMSEA [90% CI]	AIC	BIC	χ^2	*df*	*p*	CFI	RMSEA [90% CI]	AIC	BIC
Two-factor	65.75	34	<.001	.87	.10 [.06, .13]	4,566.79	4,647.55	56.25	34	.010	.88	.08 [.04, .11]	4,909.07	4,992.50
Three-factor	54.16	32	.009	.91	.08 [.04, .12]	4,555.84	4,641.81	37.84	32	.220	.97	.04 [.00, .09]	4,896.05	4,984.87
Three-factor[a]	30.80	31	.476	1.00	.00 [.00, .07]	4,539.51	4,628.09							

Note. RD = reading disorder; CFI = comparative fit index; RMSEA = root mean square error of approximation; CI = confidence interval; AIC = Akaike information criterion; BIC = Bayesian information criterion.
[a] Three-factor model with correlated residual between the object naming task and the color naming task, as indicated by modification indices.

based on nonalphanumerical stimuli. Further, the object naming task, which was administered first, might have primed the color naming: Although the object naming task consisted of black and white drawings, some children might have associated prototypical colors with the objects. As Byrne (2012) demonstrated, inclusion of correlated residuals (even if included in one of the groups only) is of no concern in testing for multigroup invariance. We thus included this additional path in the model to account for potential method effects or other sources of systematic variance across tasks.[1] By this means, the model was significantly improved as indicated by a chi-square difference test, $\Delta SB\text{-}\chi^2(1, N = 100) = 21.26$, $p < .001$ and the overall fit of this respecified model was excellent (see Table 3).

For both groups, the three-factor solution is shown in Figuree 1: The PA factor captures variance that relates to the children's awareness of phonological sound patterns and their ability to discriminate or manipulate the phonemes in spoken language. The PL factor captures the children's ability to retain acoustical information in verbal working memory. Finally, the RAN factor captures the children's ability to rapidly retrieve phonological information from long-term memory. Having established the three-factor phonological processing model as the best fitting model for each group, this model served as a baseline for the subsequent multigroup confirmatory factor analyses. Unstandardized factor loadings and item uniquenesses are presented in Table 4.

Research Question 2: Factorial and Measurement Invariance of Phonological Processing

Table 5 summarizes the results concerning group invariance. First, we tested for configural invariance and thus determined whether the number of factors as well as the pattern of free and fixed factor loadings was equal across groups. More precisely, the three-factor model that was established as a baseline model for both groups *separately* was now tested for the two groups *simultaneously*. Because, at this stage, no equality constraints are imposed, the configural model constitutes the least restrictive multigroup model. As shown in Table 5, the configural model yielded an excellent fit to the data. Thus, we can conclude that the factor pattern of phonological processing is invariant across both groups.

Second, we examined whether the linear relationships between the 10 indicators and the three underlying factors were invariant across groups (i.e., metric invariance). We therefore included equality constraints on the factor loadings in the multigroup model. Although these constraints led to a slight decrease in overall model fit, the model still provided an excellent fit to the data. Moreover, neither the chi-square difference test ($\Delta SB\text{-}\chi^2 = 11.19$, $\Delta df = 7$, $p = .130$), nor the CFI difference value (.01) suggested rejecting this model in favor of the less parsimonious configural model. Thus, invariance of factor loadings was confirmed, which indicates that the phonological processing factors are measured in the same way in both groups. In other words, the extent to which the phonological processing factors accounted for performance differences in the observed variables did not differ for children with RD versus typical learners: For both groups, a one unit change in the underlying factor scores led to the same degree of change in the observed variables.

Third, to determine whether error variances of the phonological processing model were equal across groups, equality restrictions were additionally imposed on the residual variances of the observed indicators. Again, there was a slight decrease in overall model fit. However, neither the chi-square difference test ($\Delta SB\text{-}\chi^2 = 14.99$, $\Delta df = 10$, $p = .133$), nor the CFI difference value (.01) suggested rejecting this model in favor of the less parsimonious metric model. Thus, invariance of error variances was confirmed, which suggests that for each indicator the amount of variance that is explained by the underlying factors did not differ between children with RD and typical learners. In other words, phonological processing was measured with the same amount of error in both groups.

Fourth, we examined whether the distribution of the underlying factor scores differed between groups. Therefore, equality constraints were additionally imposed on the factor variances. Again, neither the chi-square difference test ($\Delta SB\text{-}\chi^2 = 3.99$, $\Delta df = 3$, $p = .263$), nor the CFI difference value (< .01) suggested rejecting this model. Thus, we can assume children with RD and typical learners showed similar performance variation in their underlying phonological processing skills.

Last, to determine whether the relationships among the phonological processing factors were invariant across groups, we additionally imposed equality constraints on the factor covariances. Again, nested model comparison revealed a nonsignificant chi-

[1] Note that the general results of the second research question (i.e., testing for measurement invariance) were the same whether the correlated residual between object and color naming was included in the baseline model or not. In addition, we reran the analyses by using an item parcel that combined the object and color naming tasks into a single indicator in both groups rather than using correlated residuals. Again, the results of the subsequent invariance tests were the same.

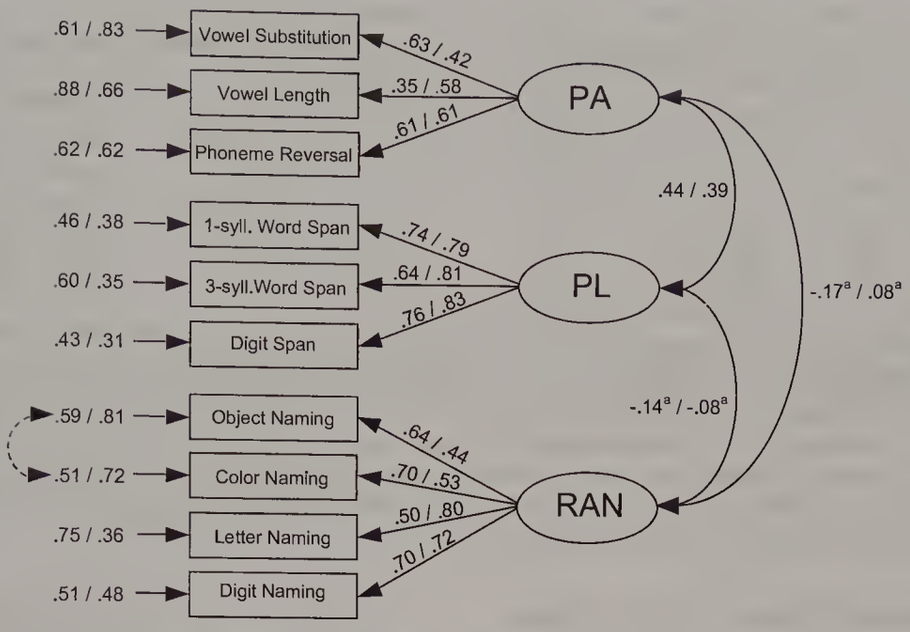


Figure 1. Three-factor oblique confirmatory model of phonological processing for children with specific reading disorder (left parameters) versus for children of the reference group (right parameters). Parameters represent standardized estimates. All loadings from the latent constructs (PA, PL, RAN) to the corresponding indicators were significant. The residual covariance between object naming and color naming (dashed line) was included in the reference group's model only. PA = phonological awareness; PL = phonological loop; RAN = rapid automatized naming; syll. = syllable. [a] Factor correlation is nonsignificant at $p < .05$.

square difference test ($\Delta SB\text{-}\chi^2 = 2.13$, $\Delta df = 3$, $p = .546$), and a CFI difference value less than .01. This result suggests that the degree to which the phonological processing factors covaried with each other did not differ between groups.

Discussion

Although children with specific RD have often been compared to typically achieving learners on various phonological processing tasks, to our knowledge no study so far has examined whether the structure of phonological processing applies to both groups of children alike. Yet, this question is important, because the medical definition of RD implies that this disorder is qualitatively distinct from normal reading (e.g., Meyer, 2000). Measurement and structural invariance is then a necessary precondition to validly compare the two groups and ensure measurement of the same construct in both groups. Therefore, this study examined the invariance of phonological processing between children with RD and their typically achieving peers. To this end, 109 third graders with RD and 100 third graders without any learning problems completed a comprehensive test battery assessing PA, RAN, and PL; we then conducted invariance tests to determine and compare the structure

Table 4
Factor Loadings and Item Uniquenesses of the Phonological Processing Model

	RD group				Reference group			
	Loadings		Uniquenesses		Loadings		Uniquenesses	
Measures	Est. (*SE*)	Std. (*SE*)	Est. (*SE*)	Std. (*SE*)	Est. (*SE*)	Std. (*SE*)	Est. (*SE*)	Std. (*SE*)
Rapid automatized naming								
Object naming	1.000 (.000)	.641 (.093)	32.480 (6.983)	.590 (.119)	1.000 (.000)	.435 (.111)	45.390 (8.196)	.811 (.096)
Colour naming	1.541 (.287)	.697 (.094)	56.874 (16.601)	.514 (.131)	1.646 (.381)	.527 (.097)	74.365 (17.161)	.722 (.102)
Digit naming	1.078 (.325)	.703 (.106)	26.94 (7.862)	.506 (.149)	1.375 (.450)	.719 (.080)	18.721 (4.454)	.484 (.115)
Letter naming	.746 (.268)	.496 (.123)	38.550 (10.774)	.754 (.122)	1.680 (.546)	.799 (.088)	16.854 (6.481)	.361 (.141)
Phonological awareness								
Vowel substitution	1.000 (.000)	.629 (.115)	5.187 (1.240)	.605 (.144)	1.000 (.000)	.418 (.192)	4.251 (1.238)	.825 (.161)
Vowel length	.424 (.192)	.349 (.116)	4.398 (.636)	.878 (.081)	1.685 (1.168)	.584 (.170)	4.960 (1.378)	.659 (.198)
Phoneme reversal	1.379 (.479)	.614 (.123)	10.633 (2.589)	.623 (.151)	3.082 (1.339)	.614 (.141)	14.145 (3.967)	.623 (.173)
Phonological loop								
One-syllabic word span	1.000 (.000)	.737 (.073)	.154 (.034)	.456 (.108)	1.000 (.000)	.786 (.052)	.158 (.029)	.383 (.081)
Three-syllabic word span	.544 (.133)	.637 (.082)	.080 (.013)	.595 (.105)	.706 (.109)	.808 (.050)	.068 (.015)	.347 (.080)
Digit span	1.010 (.173)	.755 (.070)	.141 (.033)	.430 (.106)	.985 (.131)	.828 (.048)	.114 (.027)	.314 (.079)

Note. RD = reading disorder; Est. = unstandardized parameter estimates; *SE* = standard error of the estimate; Std. = standardized parameter estimates.

Table 5
Factorial and Measurement Invariance of Phonological Processing: Fit Indices for Nested Model Comparison

Model	χ^2	*df*	p_1	Compared with	ΔSB-χ^2	Δdf	p_2	CFI	ΔCFI	RMSEA [90% CI]	AIC	BIC
Model 1	68.62	63	.293					.98		.03 [.00, .07]	9,435.57	9,659.50
Model 2	80.39	70	.186	Model 1	11.19	7	.130	.97	.01	.04 [.00, .07]	9,433.88	9,634.42
Model 3	96.13	80	.106	Model 2	14.99	10	.133	.96	.01	.04 [.00, .07]	9,431.65	9,598.76
Model 4	100.13	83	.097	Model 3	3.99	3	.263	.96	.00	.04 [.00, .07]	9,429.77	9,586.87
Model 5	102.17	86	.113	Model 4	2.13	3	.546	.96	.00	.04 [.00, .07]	9,426.00	9,573.06

Note. p_1 = probability value of model fit; SB = Satorra–Bentler; p_2 = probability value obtained in the SB chi-square difference test; CFI = comparative fit index; RMSEA = root mean square error of approximation; AIC = Akaike information criterion; BIC = Bayesian information criterion; Model 1 = configural invariance (all parameters estimated freely); Model 2 = metric invariance (factor loadings constrained equal); Model 3 = strict invariance (factor loadings and residual variances constrained equal); Model 4 = invariance of factor variances (factor loadings, residual variances, and factor variances constrained equal); Model 5 = invariance of factor covariances (factor loadings, residual variances, factor variances and factor covariances constrained equal). All models are estimated with the robust maximum likelihood estimator. In an additional analysis, Model 3 to 5 were compared with the baseline model (Model 1) rather than being tested against the respective previous model. Even under this stricter form of nested model comparison, all the resulting difference tests were nonsignificant—further confirming that measurement invariance holds across the two groups.

of these skills across both groups. The study yielded three central findings.

First, in both ability groups the model that fitted the data best was a three-factor oblique model with separate factors for PA, PL, and RAN. That is, children's phonological processing involves distinct abilities for the phonetic analysis, the short-term storage and the long-term retrieval of oral language. Our finding that a tripartite structure underlies phonological processing in normally achieving schoolchildren replicates previous studies in the field (e.g., Sprugevica & Høien, 2004), but it also adds to existing research by demonstrating that Wagner and Torgesen's (1987) phonological processing model also applies to children whose language is German, an orthography with transparent grapheme-to-phoneme correspondences. This is further evidence for the notion that the general nature of phonological processing is relatively universal, although differences across orthographies may exist in the *relative* contribution of those skills to emergent literacy (e.g., Smythe et al., 2008; Vaessen et al., 2010). Further and more importantly, we demonstrated that the tripartite structure of phonological processing also applies to children with RD: The structural organization of their phonological processing skills is as differentiated and developed as in typical learners, albeit at a less efficient level. Overall, this finding is empirical evidence for the common but so far untested assumption that phonological deficits in children with RD reflect functional rather than structural deficits. This finding challenges the medical definition of RD, which implies that children with IQ-discrepant reading problems reflect a qualitatively distinct group (e.g., Meyer, 2000). Rather, our results support a dimensional conceptualization of RD—at least with respect to underlying phonological processing: Deficits in PA, RAN, and PL seem to constitute the lower end on the population continuum rather than being a manifestation of an entirely different or less differentiated phonological processing structure.

Second, we found evidence of measurement invariance. That is, group membership did not moderate the relations between observed test scores and underlying constructs. In particular, we tested the covariance structure of the phonological processing model and found that both the measurement parameters (i.e., the factor loadings and residual variances) as well as the structural parameters (i.e., the factor variances and covariances) were invariant across children with RD and typical learners. The invariance of factor loadings implies that the phonological constructs are measured in the same way in both groups. Further, the invariance of residual variances indicates that phonological processing skills are measured with the same amount of error in both groups. In other words, there is no evidence that aspects influencing reliability such as motivation loss or floor effects are a greater issue in children with RD than in typically achieving children. Invariance of factor variances further suggests that the distribution of underlying phonological processing components did not differ between the two groups. That is, the variance of the phonological factors is neither reduced nor increased in children with RD as opposed to typical learners. Last, invariance of covariances means that the degree to which the phonological processing factors are related to each other did not differ across groups. These results have implications with respect to previous empirical studies: Traditionally, the vast majority of studies on RD comprises small to medium sample sizes due to the enormous efforts that need to be invested in recruiting participants. As a consequence, it is usually not possible to analyze the data with latent models. Group comparisons are therefore usually performed on the manifest level, which, of course, requires strict assumptions about the underlying data. Unlike in latent models, these assumptions usually cannot be checked in manifest analyses. That is, the researchers must simply assume that invariance holds and that group comparisons are thus meaningful. Hence, construct validation studies like the one presented here are of crucial importance as they provide empirical evidence for the assumption of group invariance and thereby underpin the interpretations that are drawn from manifest studies.

Finally, the correlations among the three latent factors are noteworthy. Latent correlations are often referred to as true relationships, because measurement error and task-specific effects are controlled. In both reading groups, the correlations between PA and PL were in the medium range, whereas the respective correlations with RAN were nonsignificant. This correlation pattern corresponds to previous studies (e.g., Norton & Wolf, 2012) and is in line with Wagner and Torgesen's (1987) theoretical assumption that PA, PL, and RAN each represent different facets of phonological information processing. They may therefore make quite specific and unique contributions to developing literacy skills. Consequently, all three phonological processing skills should be targeted in individual diagnostics, because profiles of phonological

strength and weaknesses may vary among subgroups of children with RD. Empirically, this idea is supported by studies which found PA and PL to be more related to reading accuracy or spelling, and RAN to be more related to reading speed (e.g., Ennemoser, Marx, Weber, & Schneider, 2012; Moll, Ramus, et al., 2014). Moreover, this correlation pattern may be taken as evidence to suggest that PA and PL on the one hand and RAN on the other hand represent independent causes for poor literacy skills. This idea is partly expressed in the double deficit hypothesis (Bowers & Wolf, 1999).

In a similar vein, there is some debate in the literature as to whether the low correlations with lexical access stem at least partially from a speed accuracy confound, because RAN but not PA and PL are measured with speeded items (cf. Wagner et al., 1993). In fact, using a speeded measure for their PA tasks, Vaessen, Gerretsen, and Blomert (2009) demonstrated that only the speeded but not the conventional PA subtests were significantly related to RAN.

Should We Move Toward a Dimensional Conceptualization of Reading Disorder?

Current ICD-10 (WHO, 2011) classification of RD is categorical in nature and expresses the idea that RD is qualitatively distinct from typical reading. Although this distinctiveness hypothesis has been criticized for years (e.g., Siegel, 1989), the surrounding debate is currently being boosted by the increased acknowledgment of the shortcomings associated with the categorical conceptualization.

Using a construct validation approach, our study provides further support for a dimensional conceptualization of RD, as we did not find any structural differences with respect to the phonological core deficit of RD. Likewise, the distinctiveness hypothesis is challenged by a range of studies comparing RD to other forms of poor reading: For instance, despite intensive research over the past decades, there is no convincing evidence that children with RD differ from non IQ-discrepant poor readers with respect to the symptoms (Hoskyn & Swanson, 2000; Stuebing et al., 2002), the cognitive causes (e.g., Jiménez, Siegel, & López, 2003; Maehler & Schuchardt, 2011; Stuebing et al., 2002; Toth & Siegel, 1994), the genetic or neuroanatomical correlates (see Stanovich, 1994b, for a review), the response to intervention (Stage, Abbott, Jenkins, & Berninger, 2003) or the general course of their reading development (e.g., Flowers, Meyer, Lovato, Wood, & Felton, 2000; O'Malley, Francis, Foorman, Fletcher, & Swank, 2002). Together, those findings are commonly interpreted as evidence for the low validity of current IQ discrepancy models in particular and the categorical approach in general (e.g., Siegel, 1992; Stanovich, 1994b; Stuebing et al., 2002).

In addition, a range of methodological papers have pointed to the negative measurement issues that result from the dichotomization of continuously distributed reading and IQ scores. For example, the arbitrary nature of the thresholds as well as problems due to regression to the mean are well-recognized phenomena (e.g., Francis et al., 2005; Sternberg & Grigorenko, 2002). The low diagnostic stability of current RD classification further emphasizes the limitations of the categorical approach (e.g., Brown Waesche, Schatschneider, Maner, Ahmed, & Wagner, 2011; Francis et al., 2005; Schatschneider, Wagner, Hart, & Tighe, 2016). Moreover, splitting the reading and IQ continuum produces an information loss within the established categories and may also create statistical artifacts and may reduce statistical power (e.g., Cohen, 1983; McCallum, Zhang, Preacher, & Rucker, 2002; Maxwell & Delaney, 1993).

To summarize, current research points to the conclusion that a dimensional conceptualization of RD might be more appropriate than a strictly categorical definition. Particularly, by acknowledging the quantitative nature of RD we might be able to improve diagnostic validity. Further, there is justified hope that a dimensional approach would significantly enhance diagnostic reliability, which would help to overcome methodological shortcomings of current diagnostics. However, further studies would still have to determine the diagnostic criteria for a dimensional assessment of RD. Overall, striving for a more valid and reliable classification of RD remains a major challenge and is of utmost importance for the improvement of intervention planning for affected children.

Limitations of the Study and Directions for Further Studies

Although our study contributes to the literature on phonological processing in RD, there are some limitations worthwhile to be considered in further research. First, although our RD group was carefully selected as part of an extreme group design and although our cut-off scores followed the diagnostic guidelines commonly used in German educational practice, thresholds to define RD are to some extent arbitrary. It might thus be possible that the tripartite structure does not hold in children who experience even more severe learning problems than the children participating in our study. Further, RD is typically conceptualized categorically. However, there is no international agreement on how this category is defined: some countries such as Germany adhere to ICD-10 (WHO, 2011), others to the *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; *DSM-5*; APA, 2013); there are also country-specific regulations. That is, countries differ in their concept of an RD, and consequently different categorical groups are formed. Against this backdrop, it has to be kept in mind that results from our study may not necessarily transfer to countries in which other diagnostic criteria are applied in RD diagnostics.

Second, our analysis is based on a limited number of tasks. For instance, we assessed PA only on the phoneme level and here mainly with vowel sounds. Future studies may therefore also incorporate subtests that refer to rhymes and syllables as well as subtests that emphasize consonant sounds. In addition, some of the subtests measuring the same underlying construct showed only low correlations and in line with this finding some of the standardized factor loadings were only in the moderate range with values of .35 to .44. Although all these factor loadings were significant, we suggest that future studies should replicate our study with measures that are related to each other to a greater extent. Likewise, the sample size of 209 children is relatively small for a multigroup analysis of invariance testing. Small samples may reduce the power to differentiate between competing models and may also produce parameter estimates with large confidence intervals. Although similar sample sizes were also used in previous studies examining measurement invariance in child mental disorders (e.g., attention-deficit/hyperactivity disorder: Karalunas, Bierman, & Huang-Pollock, 2016; intellectual disabilities: Marsh,

Tracey, & Craven, 2006; learning disabilities: Schuchardt et al., 2008), future studies should replicate the present findings based on larger sample sizes.

Third, we examined the structural organization of phonological processing at a relatively late point in children's phonological development. The rationale for investigating the nature of phonological skills among third graders was that this is the age group for which RD is most frequently diagnosed (cf. Hasselhorn & Schuchardt, 2006). Accordingly, a vast amount of studies conducted in the field targets this age group when analyzing performance differences between children with RD and typical learners, which further justifies choosing this age group as a starting point for empirical examinations of invariance. Nevertheless, we suggest that it could be worthwhile to perform a similar analysis with kindergarten children *at risk* of RD: Possibly, structural differences in phonological processing might exist at an earlier point in children's development. Likewise, it remains to be seen whether invariance would also hold longitudinally across grade levels: Because phonological processing is reciprocally related to emergent literacy skills (e.g., Burgess & Lonigan, 1998; Chow, McBride-Chang, & Burgess, 2005), it cannot be taken for granted that the structure found in one grade level automatically transfers to another grade level. Thus, longitudinal studies would help determine the structural *development* as well as the *longitudinal* invariance of phonological processing in children with RD during the first years of formal reading instruction and beyond.

Last but not least, future studies could further expand our theoretical understanding concerning the structure of phonological processing by comparing different subgroups of reading disability. For instance, it would be of interest whether children whose reading problems are manifest mainly in the domain of word decoding (referred to as *dyslexia* in the Simple View of Reading; Gough & Tunmer, 1986) show the same or a different phonological processing structure as children whose problems concentrate mainly on reading comprehension. Given the current debate as to whether these different phenotypes of RD are associated with the same or distinct cognitive deficits (see Snowling & Hulme, 2012, for a review), those studies would be highly informative with respect to the question whether we are dealing with qualitatively different subgroups or whether the structure of phonological processing is transferable to different forms of reading disability.

Implications for Educational Practice

At least two educational implications can be drawn from our study, the first of which refers to diagnostics: When it comes to diagnosing RD or to identifying at-risk children, an assessment of phonological processing is often part of the diagnostic procedure to better understand the child's strengths and weaknesses in the phonological domain. Our finding that common measures of phonological processing work equally well across typical learners and children with RD implies that the results obtained from such a diagnostic assessment have the same underlying meaning in both groups and can thus be interpreted in the same way. If, however, the measurement instruments were not invariant, the test results would imply something different with respect to the underlying phonological constructs depending on the group a child is allocated to. Moreover, as Millsap and Kwok (2004) demonstrated, diagnostic decision-making may even be biased toward one of the groups: Specifically, missing group invariance may negatively influence the sensitivity of the instrument and may result in different selection and error rates depending on the subgroup. It is thus a crucial aspect of test fairness to know whether or not a measurement instrument works equally well across the subgroups for which it is used.

Finally, given the important role of phonological processing for emergent literacy, there is growing interest in fostering these skills through intervention—an approach that has as yet proven particularly effective for PA. As regards those phonological trainings, our findings may hold specific expectations with respect to cognitive transfer. In particular, the factor structure and interrelations we found may suggest that a training of PA, for instance, may also foster phonetic recoding in the PL (and vice versa), whereas cognitive transfer to RAN does not seem very promising. Likewise, it is unlikely to expect a training of RAN to transfer to PA and PL. Although the vast majority of intervention studies have not targeted those potential transfer effects, the few existing studies provide some support for this assumption. For instance, Regtvoort and van der Leij (2007) did not find a PA training to transfer to lexical access. In contrast, there is evidence that a training of PA may booster the storage capacity of the PL (Gillam, Kleeck, & Hoffman, 2006).

To summarize, this study contributed to our theoretical understanding of phonological processing by demonstrating factorial and measurement invariance of phonological processing skills between children with RD and typical learners. In so doing, the present study closed a remaining significant gap in the RD literature.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Anthony, J. L., & Lonigan, C. J. (2004). The nature of phonological awareness: Converging evidence from four studies of preschool and early grade school children. *Journal of Educational Psychology, 96,* 43–55. http://dx.doi.org/10.1037/0022-0663.96.1.43

Anthony, J. L., Williams, J. M., McDonald, R., Corbitt-Shindler, D., Carlson, C. D., & Francis, D. J. (2006). Phonological processing and emergent literacy in Spanish-speaking preschool children. *Annals of Dyslexia, 56,* 239–270. http://dx.doi.org/10.1007/s11881-006-0011-5

Anthony, J. L., Williams, J. M., McDonald, R., & Francis, D. J. (2007). Phonological processing and emergent literacy in younger and older preschool children. *Annals of Dyslexia, 57,* 113–137. http://dx.doi.org/10.1007/s11881-007-0008-8

Archibald, L. M. D., & Gathercole, S. E. (2007). Nonword repetition and serial recall: Equivalent measures of verbal short-term memory? *Applied Psycholinguistics, 28,* 587–606. http://dx.doi.org/10.1017/S0142716407070324

Aro, M., & Wimmer, H. (2003). Learning to read: English in comparison to six more regular orthographies. *Applied Psycholinguistics, 24,* 621–635. http://dx.doi.org/10.1017/S0142716403000316

Babayiğit, S., & Stainthorp, R. (2011). Modeling the relationships between cognitive-linguistic skills and literacy skills: New insights from a transparent orthography. *Journal of Educational Psychology, 103,* 169–189. http://dx.doi.org/10.1037/a0021671

Baddeley, A. D. (1986). *Working memory.* Oxford, United Kingdom: Oxford University Press.

Birkel, P. (2007). *Weingartener Grundwortschatz Rechtschreib-Test für zweite und dritte Klassen—WRT 2* [Weingarten's spelling test of basic vocabulary for second and third grade]. Göttingen, Germany: Hogrefe.

Blomert, L., & Vaessen, A. (2008). *3DM differentiaal diagnostiek van dyslexie: Een cognitieve analyse van lezen en spellen* [3DM Differential diagnosis of dyslexia: A cognitive analysis of reading and spelling]. Amsterdam, the Netherlands: Boom.

Boscardin, C. K., Muthén, B., Francis, D. J., & Baker, E. L. (2008). Early identification of reading difficulties using heterogeneous developmental trajectories. *Journal of Educational Psychology, 100,* 192–208. http://dx.doi.org/10.1037/0022-0663.100.1.192

Bowers, P. G., & Wolf, M. (1999). The double-deficit hypothesis for the developmental dyslexias. *Journal of Educational Psychology, 91,* 415–438. http://dx.doi.org/10.1037/0022-0663.91.3.415

Bowey, J. A. (1997). What does nonword repetition measure? A reply to Gathercole and Baddeley. *Journal of Experimental Child Psychology, 67,* 295–301. http://dx.doi.org/10.1006/jecp.1997.2408

Bradley, L., & Byrant, P. E. (1985). *Rhyme and reason in reading and spelling.* Ann Arbor, MI: University of Michigan Press.

Branum-Martin, L., Fletcher, J. M., & Stuebing, K. K. (2013). Classification and identification of reading and math disabilities: The special case of comorbidity. *Journal of Learning Disabilities, 46,* 490–499.

Brown Waesche, J. S., Schatschneider, C., Maner, J. K., Ahmed, Y., & Wagner, R. K. (2011). Examining agreement and longitudinal stability among traditional and RTI-based definitions of reading disability using the affected-status agreement statistic. *Journal of Learning Disabilities, 44,* 296–307. http://dx.doi.org/10.1177/0022219410392048

Burgess, S. R., & Lonigan, C. J. (1998). Bidirectional relations of phonological sensitivity and prereading abilities: Evidence from a preschool sample. *Journal of Experimental Child Psychology, 70,* 117–141. http://dx.doi.org/10.1006/jecp.1998.2450

Byrne, B. M. (2012). *Structural equation modeling with Mplus.* New York, NY: Routledge.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York, NY: Cambridge University Press.

Cattell, R. B., Weiβ, R. H., & Osterland, J. (1997). *Grundintelligenztest Skala 1 – CFT 1* [Basic intelligence test scale 1] (5th ed.). Göttingen, Germany: Hogrefe.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233–255. http://dx.doi.org/10.1207/S15328007SEM0902_5

Chow, B. W.-Y., McBride-Chang, C., & Burgess, S. (2005). Phonological processing skills and early reading abilities in Hong Kong Chinese kindergarteners learning to read English as a second language. *Journal of Educational Psychology, 97,* 81–87. http://dx.doi.org/10.1037/0022-0663.97.1.81

Coghill, D., & Sonuga-Barke, E. J. S. (2012). Annual research review: Categories versus dimensions in the classification and conceptualisation of child and adolescent mental disorders—Implications of recent empirical study. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 53,* 469–489. http://dx.doi.org/10.1111/j.1469-7610.2011.02511.x

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249–253. http://dx.doi.org/10.1177/014662168300700301

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Academic Press.

Compton, D. L., Fuchs, L. S., Fuchs, D., Lambert, W., & Hamlett, C. (2012). The cognitive and academic profiles of reading and mathematics learning disabilities. *Journal of Learning Disabilities, 45,* 79–95. http://dx.doi.org/10.1177/0022219410393012

de Jong, P. F., & van der Leij, A. (1999). Specific contributions of phonological abilities to early reading acquisition: Results from a Dutch latent variable longitudinal study. *Journal of Educational Psychology, 91,* 450–476. http://dx.doi.org/10.1037/0022-0663.91.3.450

Denckla, M. B., & Rudel, R. G. (1976). Naming of object-drawings by dyslexic and other learning disabled children. *Brain and Language, 3,* 1–15. http://dx.doi.org/10.1016/0093-934X(76)90001-8

Dirks, E., Spyer, G., van Lieshout, E. C. D. M., & de Sonneville, L. (2008). Prevalence of combined reading and arithmetic disabilities. *Journal of Learning Disabilities, 41,* 460–473. http://dx.doi.org/10.1177/0022219408321128

Elliott, J. G., & Grigorenko, E. L. (2014). *The dyslexia debate.* New York, NY: Cambridge University Press.

Ennemoser, M., Marx, P., Weber, J., & Schneider, W. (2012). Spezifische Vorläuferfertigkeiten der Lesegeschwindigkeit, des Leseverständnisses und des Rechtschreibens: Evidenz aus zwei Längsschnittstudien vom Kindergarten bis zur 4 Klasse [Specific precursors of decoding speed, reading comprehension, and spelling: Evidence from two longitudinal studies from kindergarten to Grade 4]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 44,* 53–67. http://dx.doi.org/10.1026/0049-8637/a000057

Ferrer, E., Shaywitz, B. A., Holahan, J. M., Marchione, K., & Shaywitz, S. E. (2010). Uncoupling of reading and IQ over time: Empirical evidence for a definition of dyslexia. *Psychological Science, 21,* 93–101. http://dx.doi.org/10.1177/0956797609354084

Fischbach, A., Schuchardt, K., Brandenburg, J., Klesczewski, J., Balke-Melcher, C., Schmidt, C., . . . Hasselhorn, M. (2013). Prävalenz von Lernschwächen und Lernstörungen: Zur Bedeutung der Diagnosekriterien [Prevalence of poor learners and children with learning disorders: Investigating the role of diagnostic criteria]. *Lernen und Lernstörungen, 2,* 65–76. http://dx.doi.org/10.1024/2235-0977/a000035

Flowers, L., Meyer, M., Lovato, J., Wood, F., & Felton, R. (2000). Does third grade discrepancy status predict the course of reading development? *Annals of Dyslexia, 50,* 49–71.

Francis, D. J., Fletcher, J. M., Stuebing, K. K., Lyon, G. R., Shaywitz, B. A., & Shaywitz, S. E. (2005). Psychometric approaches to the identification of LD: IQ and achievement scores are not sufficient. *Journal of Learning Disabilities, 38,* 98–108. http://dx.doi.org/10.1177/00222194050380020101

Furnes, B., & Samuelsson, S. (2011). Phonological awareness and rapid automatized naming predicting early development in reading and spelling: Results from a cross-linguistic longitudinal study. *Learning and Individual Differences, 21,* 85–95. http://dx.doi.org/10.1016/j.lindif.2010.10.005

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics, 27,* 513–543.

Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language.* Hove, United Kingdom: Erlbaum.

Gillam, R. B., Kleeck, A. V., & Hoffman, L. M. (2006). Training in phonological awareness generalizes to phonological working memory: A preliminary investigation. *The Journal of Speech and Language Pathology, Applied Behavior Analysis, 1,* 228–243. http://dx.doi.org/10.1037/h0100201

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading and reading disability. *Remedial and Special Education, 7,* 6–10. http://dx.doi.org/10.1177/074193258600700104

Hasselhorn, M., Grube, D., & Mähler, C. (2000). Theoretisches Rahmenmodell für ein Diagnostikum zur differentiellen Funktionsanalyse des phonologischen Arbeitsgedächtnisses [Theoretical framework for diagnosing the differential functional analysis of phonological working memory]. In M. Hasselhorn, W. Schneider, & H. Marx (Eds.), *Diagnostik von Lese-Rechtschreibschwierigkeiten (Test und Trends N. F., Bd. 1. Jahrbuch der pädagogisch-psychologischen Diagnostik*; pp. 167–182). Göttingen, Germany: Hogrefe.

Hasselhorn, M., Mähler, C., & Grube, D. (2008). Lernstörungen in Teilleistungsbereichen [Learning disabilities in specific domains]. In R. Oerter & L. Montada (Eds.), *Entwicklungspsychologie* (pp. 769–778). Weinheim, Germany: Psychologie Verlagsunion.

Hasselhorn, M., & Schuchardt, K. (2006). Lernstörungen: Eine kritische Skizze zur Epidemiologie [Learning disabilities: A critical sketch on epidemiology]. *Kindheit und Entwicklung, 15,* 208–215. http://dx.doi.org/10.1026/0942-5403.15.4.208

Hasselhorn, M., Schumann-Hengsteler, R., Grube, D., König, J., Mähler, C., Schmid, I., . . . Zoelch, C. (2012). *Arbeitsgedächtnistestbatterie für Kinder von 5 bis 12 Jahren (AGTB 5–12)* [Working memory test battery for children aged five to twelve years]. Göttingen, Germany: Hogrefe.

Hoskyn, M., & Swanson, H. L. (2000). Cognitive processing of low achievers and children with reading disabilities: A selective meta-analytic review of the published literature. *School Psychology Review, 29,* 102–119.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Jiménez, J. E., Siegel, L. S., & López, M. R. (2003). The relationship between IQ and reading disabilities in English-speaking Canadian and Spanish children. *Journal of Learning Disabilities, 36,* 15–23. http://dx.doi.org/10.1177/00222194030360010301

Karalunas, S. L., Bierman, K. L., & Huang-Pollock, C. L. (2016). Test–retest reliability and measurement invariance of executive function tasks in young children with and without ADHD. *Journal of Attention Disorders.* Advance online publication. http://dx.doi.org/10.1177/1087054715627488

Kirby, J. R., Georgiou, G. K., Martinussen, R., Parrila, R., Bowers, P., & Landerl, K. (2010). Naming speed and reading: From prediction to instruction. *Reading Research Quarterly, 45,* 341–362. http://dx.doi.org/10.1598/RRQ.45.3.4

Klicpera, C., Schabmann, A., & Gasteiger-Klicpera, B. (2010). *Legasthenie – LRS: Modelle, diagnose, therapie und förderung* [Dyslexia: Models, diagnosis, therapy and intervention]. München, Germany: Reinhardt.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford Press.

Krajewski, K., Liehm, S., & Schneider, W. (2004). *Deutscher Mathematiktest für zweite Klassen (DEMAT 2+)* [German test of mathematical skills for second grades]. Göttingen, Germany: Beltz.

Kudo, M. F., Lussier, C. M., & Swanson, H. L. (2015). Reading disabilities in children: A selective meta-analysis of the cognitive literature. *Research in Developmental Disabilities, 40,* 51–62. http://dx.doi.org/10.1016/j.ridd.2015.01.002

Lambrecht Smith, S., Scott, K. A., Roberts, J., & Locke, J. L. (2008). Disabled readers' performance on tasks of phonological processing, rapid naming, and letter knowledge before and after kindergarten. *Learning Disabilities Research & Practice, 23,* 113–124. http://dx.doi.org/10.1111/j.1540-5826.2008.00269.x

Landerl, K. (2001). Word recognition deficits in German: More evidence from a representative sample. *Dyslexia, 7,* 183–196. http://dx.doi.org/10.1002/dys.199

Landerl, K. (2003). Categorization of vowel length in German poor spellers! An orthographically relevant phonological distinction. *Applied Psycholinguistics, 24,* 523–538. http://dx.doi.org/10.1017/S0142716403000262

Landerl, K., & Moll, K. (2010). Comorbidity of learning disorders: Prevalence and familial transmission. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 51,* 287–294. http://dx.doi.org/10.1111/j.1469-7610.2009.02164.x

Landerl, K., Wimmer, H., & Frith, U. (1997). The impact of orthographic consistency on dyslexia: A German-English comparison. *Cognition, 63,* 315–334. http://dx.doi.org/10.1016/S0010-0277(97)00005-X

Lenhard, W., & Schneider, W. (2006). *Ein Leseverständnistest für Erst- bis Sechstklässler (ELFE 1–6)* [A reading comprehension test for first to sixth graders]. Göttingen, Germany: Hogrefe.

Logan, J. A. R., Schatschneider, C., & Wagner, R. K. (2011). Rapid serial naming and reading ability: The role of lexical access. *Reading and Writing, 24,* 1–25. http://dx.doi.org/10.1007/s11145-009-9199-1

Lonigan, C. J., Anthony, J. L., Phillips, B. M., Purpura, D. J., Wilson, S. B., & McQueen, J. D. (2009). The nature of preschool phonological processing abilities and their relations to vocabulary, general cognitive abilities, and print knowledge. *Journal of Educational Psychology, 101,* 345–358. http://dx.doi.org/10.1037/a0013837

Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2003). Defining dyslexia, comorbidity, teachers' knowledge of language and reading: A definition of dyslexia. *Annals of Dyslexia, 53,* 1–14. http://dx.doi.org/10.1007/s11881-003-0001-9

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7,* 19–40. http://dx.doi.org/10.1037/1082-989X.7.1.19

Maehler, C., & Schuchardt, K. (2011). Working memory in children with learning disabilities: Rethinking the criterion of discrepancy. *International Journal of Disability Development and Education, 58,* 5–17. http://dx.doi.org/10.1080/1034912X.2011.547335

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics—Series B, 36,* 115–128.

Marsh, H. W., Tracey, D. K., & Craven, G. H. (2006). Multidimensional self-concept structure for preadolescents with mild intellectual disabilities: A hybrid multigroup–MIMC approach to factorial invariance and latent mean differences. *Educational and Psychological Measurement, 66,* 795–818. http://dx.doi.org/10.1177/0013164405285910

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Quantitative Methods in Psychology, 113,* 181–190.

Melby-Lervåg, M., Lyster, S.-A. H., & Hulme, C. (2012). Phonological skills and their role in learning to read: A meta-analytic review. *Psychological Bulletin, 138,* 322–352. http://dx.doi.org/10.1037/a0026744

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525–543. http://dx.doi.org/10.1007/BF02294825

Metsala, J. L. (1999). Young children's phonological awareness and nonword repetition as a function of vocabulary development. *Journal of Educational Psychology, 91,* 3–19. http://dx.doi.org/10.1037/0022-0663.91.1.3

Meyer, M. S. (2000). The ability–achievement discrepancy: Does it contribute to an understanding of learning disabilities? *Educational Psychology Review, 12,* 315–337. http://dx.doi.org/10.1023/A:1009070006373

Michalczyk, K., Malstädt, N., Worgt, M., Könen, T., & Hasselhorn, M. (2013). Age differences and measurement invariance of working memory in 5- to 12-year-old children. *European Journal of Psychological Assessment, 29,* 220–229. http://dx.doi.org/10.1027/1015-5759/a000149

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* New York, NY: Routledge.

Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9,* 93–115. http://dx.doi.org/10.1037/1082-989X.9.1.93

Mody, M. (2003). Phonological basis in reading disability: A review and analysis of the evidence. *Reading and Writing, 16,* 21–39. http://dx.doi.org/10.1023/A:1021741921815

Moll, K., Fussenegger, B., Willburger, E., & Landerl, K. (2009). RAN is not a measure of orthographic processing: Evidence from the asymmetric German orthography. *Scientific Studies of Reading, 13,* 1–25. http://dx.doi.org/10.1080/10888430802631684

Moll, K., Kunze, S., Neuhoff, N., Bruder, J., & Schulte-Körne, G. (2014). Specific learning disorder: Prevalence and gender differences. *PLoS ONE, 9,* e103537. http://dx.doi.org/10.1371/journal.pone.0103537

Moll, K., Ramus, F., Bartling, J., Bruder, J., Kunze, S., Neuhoff, N., . . . Landerl, K. (2014). Cognitive mechanisms underlying reading and spelling development in five European orthographies. *Learning and Instruction, 29,* 65–77. http://dx.doi.org/10.1016/j.learninstruc.2013.09.003

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Niedersächsisches Kultusministerium. (2006). *Kerncurriculum für die Grundschule: Schuljahrgänge 1–4* [Core-curriculum for elementary school: Grades 1–4]. Retrieved from http://db2.nibis.de/1db/cuvo/ausgabe/

Nikolopoulos, D., Goulandris, N., Hulme, C., & Snowling, M. J. (2006). The cognitive bases of learning to read and spell in Greek: Evidence from a longitudinal study. *Journal of Experimental Child Psychology, 94,* 1–17. http://dx.doi.org/10.1016/j.jecp.2005.11.006

Norton, E. S., & Wolf, M. (2012). Rapid automatized naming (RAN) and reading fluency: Implications for understanding and treatment of reading disabilities. *Annual Review of Psychology, 63,* 427–452. http://dx.doi.org/10.1146/annurev-psych-120710-100431

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

O'Malley, K. J., Francis, D. J., Foorman, B. R., Fletcher, J. M., & Swank, P. R. (2002). Growth in precursor and reading-related skills: Do low-achieving and IQ-discrepant readers develop differently? *Learning Disabilities Research & Practice, 17,* 19–34. http://dx.doi.org/10.1111/1540-5826.00029

Papadopoulos, T. C., Kendeou, P., & Spanoudis, G. (2012). Investigating the factor structure and measurement invariance of phonological abilities in a sufficiently transparent language. *Journal of Educational Psychology, 104,* 321–336. http://dx.doi.org/10.1037/a0026446

Papadopoulos, T. C., Spanoudis, G., & Kendeou, P. (2009). The dimensionality of phonological abilities in Greek. *Reading Research Quarterly, 44,* 127–143. http://dx.doi.org/10.1598/RRQ.44.2.2

Peterson, R. L., & Pennington, B. F. (2015). Developmental dyslexia. *Annual Review of Clinical Psychology, 11,* 283–307. http://dx.doi.org/10.1146/annurev-clinpsy-032814-112842

Regtvoort, G. F. M., & van der Leij, A. (2007). Early intervention with children of dyslexic parents: Effects of computer-based reading instruction at home on literacy acquisition. *Learning and Individual Differences, 17,* 35–53. http://dx.doi.org/10.1016/j.lindif.2007.01.005

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrica, 66,* 507–514. http://dx.doi.org/10.1007/BF02296192

Schatschneider, C., Francis, D. J., Forrman, B. R., Fletcher, J. M., & Mehta, P. (1999). The dimensionality of phonological awareness: An application of item response theory. *Journal of Educational Psychology, 91,* 439–449. http://dx.doi.org/10.1037/0022-0663.91.3.439

Schatschneider, C., Wagner, R. K., Hart, S. A., & Tighe, E. L. (2016). Using simulations to investigate the longitudinal stability of alterntive schemas for classifying and identifying children with reading disabilities. *Scientific Studies of Reading, 20,* 34–48. http://dx.doi.org/10.1080/10888438.2015.1107072

Schneider, W., & Näslund, J. C. (1993). The impact of early metalinguistic competencies and memory capacity on reading and spelling in elementary school: Results of the Munich longitudinal study on the genesis of individual competencies (LOGIC). *European Journal of Psychology of Education, 8,* 273–287. http://dx.doi.org/10.1007/BF03174082

Schuchardt, K., Roick, T., Mähler, C., & Hasselhorn, M. (2008). Unterscheidet sich die Struktur des Arbeitsgedächtnisses bei Schulkindern mit und ohne Lernstörung? [Does learning disability make a difference regarding the structure of working memory in children?]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 40,* 147–151. http://dx.doi.org/10.1026/0049-8637.40.3.147

Siegel, L. S. (1989). IQ is irrelevant to the definition of learning disabilities. *Journal of Learning Disabilities, 22,* 469–478. http://dx.doi.org/10.1177/002221948902200803

Siegel, L. S. (1992). An evaluation of the discrepancy definition of dyslexia. *Journal of Learning Disabilities, 25,* 618–629. http://dx.doi.org/10.1177/002221949202501001

Smythe, I., Everatt, J., Al-Menaye, N., He, X., Capellini, S., Gyarmathy, E., & Siegel, L. S. (2008). Predictors of word-level literacy amongst Grade 3 children in five diverse languages. *Dyslexia, 14,* 170–187. http://dx.doi.org/10.1002/dys.369

Snowling, M. J., & Hulme, C. (2012). Annual research review: The nature and classification of reading disorders—A commentary on proposals for *DSM–5. Journal of Child Psychology and Psychiatry, and Allied Disciplines, 53,* 593–607. http://dx.doi.org/10.1111/j.1469-7610.2011.02495.x

Sprugevica, I., & Høien, T. (2004). Relations between enabling skills and reading comprehension: A follow-up study of Latvian students from first to second grade. *Scandinavian Journal of Psychology, 45,* 115–122. http://dx.doi.org/10.1111/j.1467-9450.2004.00386.x

Stage, S. A., Abbott, R. D., Jenkins, J. R., & Berninger, V. W. (2003). Predicting response to early reading intervention from verbal IQ, reading-related language abilities, attention ratings, and verbal IQ-word reading discrepancy: Failure to validate discrepancy method. *Journal of Learning Disabilities, 36,* 24–33. http://dx.doi.org/10.1177/002221940303600010401

Stanovich, K. E. (1988). Explaining the differences between the dyslexic and the garden-variety poor reader: The phonological-core variable-difference model. *Journal of Learning Disabilities, 21,* 590–604. http://dx.doi.org/10.1177/002221948802101003

Stanovich, K. E. (1994a). Annotation: Does dyslexia exist? *Child Psychology & Psychiatry & Applied Disciplines, 35,* 579–595. http://dx.doi.org/10.1111/j.1469-7610.1994.tb01208.x

Stanovich, K. E. (1994b). Are discrepancy-based definitions of dyslexia empirically defensible? In K. P. van den Bos, L. S. Siegel, D. J. Bakker, & D. L. Share (Eds.), *Current directions in dyslexia research* (pp. 15–30). Lisse, the Netherlands: Swets & Zeitlinger.

Sternberg, R. J., & Grigorenko, E. L. (2002). Difference scores in the identification of children with learning disabilities—It's time to use a different method. *Journal of School Psychology, 40,* 65–83. http://dx.doi.org/10.1016/S0022-4405(01)00094-2

Stock, C., Marx, P., & Schneider, W. (2003). *Basiskompetenzen für Lese-Rechtschreibleistungen: Ein Test zur Erfassung der phonologischen Bewusstheit vom ersten bis vierten Schuljahr (BAKO 1–4)* [Basic competencies of reading and spelling skills: A test assessing phonological awareness from first to fourth grade]. Göttingen, Germany: Beltz.

Strehlow, U., & Haffner, J. (2002). Definitionsmöglichkeiten und sich daraus ergebende Häufigkeit der umschriebenen Lese- bzw. Rechtschreibstörung – theoretische Überlegungen und empirische Befunde an einer repräsentativen Stichprobe junger Erwachsener [Alternative definitions and the resulting prevalence rates of spelling disorder – Theoretical considerations and empirical findings in an epidemiological sample of adolescents and young adults]. *Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie, 30,* 113–126. http://dx.doi.org/10.1024//1422-4917.30.2.113

Stuebing, K. K., Fletcher, J. M., LeDoux, J. M., Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2002). Validity of IQ-discrepancy classifications of reading disabilities: A meta-analysis. *American Educational Research Journal, 39,* 469–518. http://dx.doi.org/10.3102/00028312039002469

Swanson, H. L. (2012). Adults with reading disabilities: Converting a meta-analysis to practice. *Journal of Learning Disabilities, 45,* 17–30. http://dx.doi.org/10.1177/0022219411426856

Swanson, H. L., & Hsieh, C.-J. (2009). Reading disabilities in adults: A selective meta-analysis of the literature. *Review of Educational Research, 79,* 1362–1390. http://dx.doi.org/10.3102/0034654309350931

Swanson, H. L., Trainin, G., Necoechea, D. M., & Hammill, D. D. (2003). Rapid naming, phonological awareness, and reading: A meta-analysis of the correlation evidence. *Review of Educational Research, 73,* 407–440. http://dx.doi.org/10.3102/00346543073004407

Torppa, M., Parrila, R., Niemi, P., Lerkkanen, M.-K., Poikkeus, A.-M., & Nurmi, J.-E. (2013). The double deficit hypothesis in the transparent Finnish orthography: A longitudinal study from kindergarten to Grade 2. *Reading and Writing, 26,* 1352–1380. http://dx.doi.org/10.1007/s11145-012-9423-2

Toth, G., & Siegel, L. S. (1994). A critical evaluation of the IQ-achievement discrepancy based definitions of dyslexia. In K. P. van den Bos, L. S. Siegel, D. J. Bakker, & D. L. Share (Eds.), *Current directions in dyslexia research* (pp. 45–70). Lisse, the Netherlands: Swets & Zeitlinger.

Vaessen, A., Bertrand, D., Tóth, D., Csépe, V., Faísca, L., Reis, A., & Blomert, L. (2010). Cognitive development of fluent word reading does not qualitatively differ between transparent and opaque orthographies. *Journal of Educational Psychology, 102,* 827–842. http://dx.doi.org/10.1037/a0019465

Vaessen, A., Gerretsen, P., & Blomert, L. (2009). Naming problems do not reflect a second independent core deficit in dyslexia: Double deficits explored. *Journal of Experimental Child Psychology, 103,* 202–221. http://dx.doi.org/10.1016/j.jecp.2008.12.004

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–70. http://dx.doi.org/10.1177/109442810031002

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9,* 486–492. http://dx.doi.org/10.1080/17405629.2012.686740

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 45,* 2–40. http://dx.doi.org/10.1046/j.0021-9630.2003.00305.x

Vloedgraven, J., & Verhoeven, L. (2009). The nature of phonological awareness throughout the elementary grades: An item response theory perspective. *Learning and Individual Differences, 19,* 161–169. http://dx.doi.org/10.1016/j.lindif.2008.09.005

Wagner, R. K. (1986). Phonological processing abilities and reading: Implications for disabled readers. *Journal of Learning Disabilities, 19,* 623–630. http://dx.doi.org/10.1177/002221948601901009

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin, 101,* 192–212. http://dx.doi.org/10.1037/0033-2909.101.2.192

Wagner, R. K., Torgesen, J. K., Laughon, P., Simmons, K., & Rashotte, C. A. (1993). Development of young readers' phonological processing abilities. *Journal of Educational Psychology, 85,* 83–103. http://dx.doi.org/10.1037/0022-0663.85.1.83

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1994). Development of reading-related phonological processing abilities: New evidence of bi-directional causality from a latent variable longitudinal study. *Developmental Psychology, 30,* 73–87. http://dx.doi.org/10.1037/0012-1649.30.1.73

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus.* West Sussex, United Kingdom: Wiley. http://dx.doi.org/10.1002/9781118356258

Wimmer, H., Landerl, K., Linortner, R., & Hummer, P. (1991). The relationship of phonemic awareness to reading acquisition: More consequence than precondition but still important. *Cognition, 40,* 219–249. http://dx.doi.org/10.1016/0010-0277(91)90026-Z

Wimmer, H., & Schurz, M. (2010). Dyslexia in regular orthographies: Manifestation and causation. *Dyslexia, 16,* 283–299. http://dx.doi.org/10.1002/dys.411

Wolf, M., Goldberg O'Rourke, A., Gidney, C., Lovett, M., Cirino, P., & Morris, R. (2002). The second deficit: An investigation of the independence of phonological and naming-speed deficits in developmental dyslexia. *Reading and Writing, 15,* 43–72. http://dx.doi.org/10.1023/A:1013816320290

World Health Organization. (2011). *ICD: Classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines* (10th rev. ed.). Geneva, Switzerland: Author.

Ziegler, J. C., Perry, C., Ma-Wyatt, A., Ladner, D., & Schulte-Körne, G. (2003). Developmental dyslexia in different languages: Language-specific or universal? *Journal of Experimental Child Psychology, 86,* 169–193. http://dx.doi.org/10.1016/S0022-0965(03)00139-5

Received June 8, 2015
Revision received August 26, 2016
Accepted September 19, 2016 ■

Journal of Educational Psychology
2017, Vol. 109, No. 5, 727–740

0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000166

Peer Influence on Children's Reading Skills: A Social Network Analysis of Elementary School Classrooms

North Cooc
The University of Texas at Austin

James S. Kim
Harvard University

Research has found that peers influence the academic achievement of children. However, the mechanisms through which peers matter remain underexplored. The present study examined the relationship between peers' reading skills and children's own reading skills among 4,215 total second- and third-graders in 294 classrooms across 41 schools. One innovation of the study was the use of social network analysis to directly assess who children reported talking to or seeking help from and whether children who identified peers with stronger reading skills experienced higher reading skills. The results indicated that children on average identified peers with stronger reading skills and the positive association between peer reading skills and children's own reading achievement was strongest for children with lower initial levels of reading skills. The study has implications for how teachers can leverage the advantages of peers via in-class activities.

Educational Impact and Implications Statement
This study shows that early elementary schoolchildren report identifying and interacting with peers with stronger reading skills within the same classroom. Children with low initial reading skills are more likely to identify such peers than high achieving children and those who do so experience higher reading outcomes later on. The study suggests that peer effects may occur through the peer-seeking patterns of children and the direct expertise of their peers.

Keywords: peer effects, reading skills, social networks

Supplemental materials: http://dx.doi.org/10.1037/edu0000166.supp

The influence of peers on student learning and achievement has long been an interest of educators, parents, and researchers. Classroom decisions about student grouping, whether through formal tracking or informal reading activities, are often guided by beliefs about how students interact and learn from each other (e.g., Hong, Corter, Hong, & Pelletier, 2012). Parents may also make decisions about schools in part because of beliefs about the advantages of learning from high-achieving peers, the benefits of adopting the norms established in these environments, or the value of a diverse population (Kimelberg & Billingham, 2013; Roda & Wells, 2013). These beliefs about student learning are supported in extensive empirical research showing that peer interactions and relationships are associated with a range of adolescent behaviors and long-term academic outcomes (Fujimoto, Unger, & Valente, 2012; Justice, Petscher, Schatschneider, & Mashburn, 2011; Mashburn, Justice, Downer, & Pianta, 2009; Sacerdote, 2001; Zimmer & Toma, 2000). The extent to which peers matter relative to other family and school inputs is important for educators in addressing disparities in academic achievement and designing appropriate instructional practices. Consequently, numerous studies across disciplines have focused on identifying how peer effects are distributed across schools and whether they matter more for some students than others (e.g., Hanushek, Kain, Markman, & Rivkin, 2003; Justice et al., 2011).

Although evidence suggests that peers matter in many educational contexts, few studies examine *how* peers influence one another (see review in Harris, 2010). Part of the challenge is linking theories of peer influence to empirical evidence. Different theories may yield similar or ambiguous predictions of student outcomes, making it difficult to test the mechanisms of peer influence. For instance, peers may have a direct influence through student-to-student interaction in learning groups, whereas an indirect influence may occur via specific norms within a classroom of similar peers. Studies that define peer influence broadly are less able to distinguish the two, which has implications for how to tailor classroom practice. A related challenge is how studies op-

This article was published Online First December 12, 2016.

North Cooc, College of Education, The University of Texas at Austin; James S. Kim, Graduate School of Education, Harvard University.

This study was made possible with an Investing in Innovation Fund (i3) grant from the U.S. Department of Education (PR/Award U396B100195). However, the contents of this article do not represent the policy of the U.S. Department, and the content is solely the responsibility of the authors. We thank George Farkas, Ha Yeon Kim, and Young-Suk Kim for comments on earlier versions of this article.

Correspondence concerning this article should be addressed to North Cooc, College of Education, The University of Texas at Austin, 1912 Speedway, Austin, TX 78712. E-mail: ncooc@austin.utexas.edu

erationalize the concept of peer influence with empirical data. Different measures of peer effect may lead to different findings across studies that can affect policy decisions. Both challenges—theory and measurement—are important to address for educators to develop more informed interventions that leverage the benefits of peers (Mashburn et al., 2009).

In this study, we address the theoretical and measurement challenges in the peer influence literature by using a social network analysis approach that more directly identifies who children report talking to and seeking help from in classrooms and assesses whether the average achievement of these peer groups predicts individual achievement. We also replicate and extend previous studies in several ways. Whereas many other studies have used smaller classroom samples to examine peer influence (e.g., Delay, Hanish, Martin, & Fabes, 2016; Justice et al., 2011), we surveyed more than 4,200 children about their peer relationships. This replication, which to our knowledge is the largest to examine peer influence on young children, improves on the generalizability and stability of prior findings. We expanded on the literature by using improved measures of who children report interacting with and the characteristics of those peers to explore the mechanisms through which peer effects may occur. This is important because knowing that children in classrooms with high achievers on average tend to perform better than in classroom with low achievers, as previous research indicate (e.g., Hanushek et al., 2003), is less useful to teachers on a day-to-day basis than understanding the relationships between peers and how students learn from each other. This study addressed whether children reported identifying peers who are strong readers within the classroom, which allowed us to analyze one potential mechanism of peer influence. The results have implications for classroom instruction decisions related to structured peer activities and class time for students to work together. Lastly, to better support struggling readers, this study analyzed whether interacting with higher achieving peers is associated with larger benefits for students who have lower initial reading achievement. Finding differential effects for struggling readers provides another way to reduce achievement gaps.

To be clear, throughout the paper and discussion of the methods and results, our use of the terms "peer influence" and "peer effects" refer to statistical associations between the achievement of children and their peers, controlling for confounding factors. We focus on describing these associations in peer relationships and acknowledge here and in the limitations section that the design of the study was not intended to allow for causal inferences about peers. Although we have made efforts to avoid the use of causal language, the terms "peer influence" and "peer effects" in this paper refer only to statistical associations in a nonexperimental study. We have retained these terms to be consistent with the research literature on this topic.

Theories of Peer Influence

Social Contagion

Studies exploring how peers influence academic outcomes or adolescent behaviors are generally grounded in an "epidemic" or "contagion" theory in which children emulate their peers (Christakis & Fowler, 2013; Jencks & Mayer, 1990). For instance, if peers are high achievers who engage in academically oriented habits like studying and homework completion, then students who interact with these peers may adopt those habits and perform better academically (Harris, 2010). The contagion model suggests that these habits maybe absorbed without explicit modeling from peers. This process also implies differential benefits whereby more disadvantaged students gain more from peers who are stronger academically (e.g., Justice, Logan, Lin, & Kaderavek, 2014), whereas peers who struggle academically may lower a student's own scores. A concern within preschool programs that enroll mostly low-income students is that children are less exposed to peers with school readiness skills who can indirectly provide academic stimulation (Schechter & Bye, 2007). Studies on children's language growth have found that children with lower initial language skills saw more gains when in classrooms with peers who had higher levels of language development (Justice et al., 2011; Mashburn et al., 2009). Related research indicates similar contagion processes occur when examining peer behaviors like smoking (Fujimoto et al., 2012). Although these studies provide support of the contagion model, it is unclear how children interact with each other beyond attending the same classroom or school.

Instrumental Model

In contrast to the social contagion models, psychologists have emphasized the mediating and active role of peer collaboration to enhance the cognitive development of lower-ability children. These theories argue that peers can provide expertise or directly model behaviors and skills that are "instrumental" to academic achievement (Harris, 2010; Ryan & Shim, 2012). Expert guidance (Vygotsky, 1978; Wood, Bruner, & Ross, 1976) and observational learning (Bandura, 1977, 1986; Schunk, 1998) are potentially two mechanisms through which peer effects operate. For example, in an experimental study involving preschool children, Azmitia (1988) found that peer collaboration promoted greater learning than independent work, particularly for novice learners. Peer effects were largely mediated through experts' guidance in supporting the learning of novices and novices' own initiative in observing, imitating, and learning from experts (Vygotsky, 1978; Ellis & Rogoff, 1986). Importantly, experimental research indicates that peer effects involve both experts' direct assistance in supporting novices' learning and novices' help-seeking behaviors in learning from experts (Bandura, 1977; Wood, 1980). Observation of peer models can raise observers' self-efficacy beliefs, which in turn may influence learning and achievement (Schunk, 1987; Schunk & Hanson, 1989). Quasi-experimental research in school and classroom contexts also indicates that lower-ability children experience larger academic gains when learning with higher-ability peers (Hanushek et al., 2003; Harris, 2010; Justice et al., 2011). Studies that define peer influence broadly, such as attending a classroom where the average student achievement level is high, are less able to highlight potential mechanisms through which peers influence an individual child's learning.

Homophily

One empirical challenge in the peer influence literature is separating effects of peers from selection bias. That is, high achieving children may simply identify or interact with peers who also have high levels of academic achievement (Altermatt & Pomerantz,

2003, 2005). Homophily is the social phenomenon and tendency of individuals to affiliate with others who share similar attributes (Kandel, 1978). Such affiliations have been observed in terms of children's externalizing behaviors (Hanish, Martin, Fabes, Leonard, & Herzog, 2005) and academic motivation (Ryan, 2001), as well as attributes like gender and race (Vu & Locke, 2014). If children affiliate with similar friends then they are likely to act or behave in similar ways, thus making it difficult to distinguish peer influence—via social contagion or instrumental effects—from homophily. Children may have higher academic achievement not because of peer effects of learning from each other but because of other unobserved traits that led them to interact. A robust literature indicates that peer effects via observational learning may depend on perceived similarity between model and observer in terms of academic competence and skills (e.g., Schunk, 1987). A key question is whether children, particularly those with lower achievement, identify and benefit from associating with peers with higher achievement. Identifying this social network pattern in classrooms would provide stronger evidence of peer influence on achievement, even if children are likely affiliating with similar peers on other dimensions.

In general, numerous theories and mechanisms are involved when examining peer influence. In studies that specifically examine language skills and other academic outcomes (e.g., Justice et al., 2011; Mashburn et al., 2009), the role of peer expertise and assistance is assumed but rarely explored. However, examining the mechanism is important if we are to understand *how* peers matters in relation to helping children develop skills like reading. For instance, whether it is sufficient for school officials to know that children are around peers who are strong readers or in a classroom with similar readers is important for instructional practice. Theories that emphasize the role of peer collaboration should ideally explore whether lower-ability readers identify higher-ability readers, and whether peer effects depend on children's individual skill.

Measuring Peer Influence

Although studies make implicit assumptions about the social contagion and instrumental models that inform how peer influence is manifested in school settings, part of the issue is how researchers measure peer influence. Measurement is important not only for accuracy but also the type of inferences that one can make about peers. In many empirical studies, particularly those using large student administrative records or secondary data, peer influence is typically operationalized as the average achievement level of students within a given classroom (e.g., Hanushek, Kain, Markman, & Rivkin, 2003; Henry & Rickman, 2007; Mashburn et al., 2009). The assumption is that a classroom of high achieving classmates contributes an overall peer effect on a given student, such as higher reading scores. The classroom average is more aligned with the social contagion model of peer influence, but it overlooks the smaller and more informal social networks that children may form with one another based on proximity or shared interests. In the case of developing reading skills, it is likely through these more proximal networks that peers may influence student achievement, especially if experts assist novices by providing corrective feedback on literacy tasks (Fuchs, Fuchs, Mathes, & Simmons, 1997; Greenwood, Delquadri, & Hall, 1989) or novices ask experts for help, or learn by observing experts (Bandura, 1977).

Examining the instrumental model of peer influence requires a different measure of peers. In sociology and public health, this approach typically involves identifying all the peers of each individual or specific peers, such those from whom one seeks advice or academic help (Scott & Carrington, 2011). Rather than using a class average, the peer influence variable is the average score of only the peers who were identified for each student (Frank et al., 2008). Although surveying young children about peer relationships can be a difficult procedure, the measure provides a more precise summary of who children interact with regularly and proximal peers. For instance, if struggling readers are able to identify peers who are strong readers and seek them out for help, then this type of social network measure is potentially more helpful in assessing how peers may influence achievement. Unfortunately, studies using the aggregate classroom or school achievement level of children are unable to examine this potential mechanism.

Why Focus on Peer Influence on Early Reading Skills?

The National Assessment of Educational Process (NAEP), a low-stakes but representative assessment of students across the United States, indicated in 2013 that only 42% of fourth graders scored at or above proficient in reading (National Center for Education Statistics, 2013). Nearly 20% of students scored below basic and can be considered struggling readers who lack the ability to comprehend written text at grade level. These trends are concerning because basic reading and literacy skills are critical for acquiring content knowledge and strongly predict future outcomes like graduation, employment, and college (Achieve Inc, 2005; Kamil, 2003; Snow & Biancarosa, 2003). Although school intervention efforts can improve children's cognitive and reading skills (see synthesis in Edmonds et al., 2009), social scientists have theorized and found that peer composition is strongly associated with individual children's reading development in the elementary grades (Entwisle & Alexander, 1994; Kindermann, 2007).

Experimental research indicates that peer collaboration is a particularly powerful and malleable factor that can impact children's early literacy development. For example, the Peer Assisted Learning Strategies (PALS) utilize the peers of struggling readers as coaches to help children acquire reading skills (Fuchs et al., 2000; Fuchs, Fuchs, & Burish, 2000). Students are put in pairs geared toward their individual needs, rather than a single teacher-directed activity that may not address the reading challenges of most children. In shifting instruction from teachers to students, these strategies can be both effective and efficient in helping struggling readers. Results from multiple replication experiments have shown that PALS improves a range of early literacy skills, including children's phonological awareness, word reading ability, and oral reading fluency outcomes (Lemons et al., 2014), which are moderately correlated with later reading comprehension outcomes (Good et al., 2011). The role of peers for struggling readers is particularly important because they appear to benefit more from exposure to peers with stronger literacy skills (Justice et al., 2011; Mashburn et al., 2009).

Although peer reading interventions like PALS require some formal structure in matching students, most center on the basic premise that peers matter and struggling readers can develop literacy skills when interacting with certain peers. In other words,

these interventions are a highly structured form of the social contagion and instrumental models of peer influence. However, the peer effects literature suggests that these processes are also like to occur naturally and informally, while still accruing benefits for students. Understanding the extent to which this happens can be helpful for teachers in designing practices that leverage the skills and advantages of peers. Furthermore, although tools like the mCLASS-DIBELS Next is used to group students with similar needs, teachers can apply the same principles to create heterogeneous groups if students learn better from peers from diverse achievement backgrounds.

Present Study

The consistent finding that peers appear to influence the academic achievement of children is not new, but the present study extends the literature in several ways. First, we assessed *how* peers may influence reading skills by examining the characteristics of peers. If peers matter because they provide access to expertise then evidence that students identify and interact with these peers can provide support for this theory of peer influence. Although students may benefit from simply learning in the same classroom as certain peers and observing similar norms, we focused on their peer interactions and own agency in developing reading skills.

Our second contribution focuses on using more direct measures of peer influence when examining its effects on academic achievement. Prior studies often operationalize peer influence as the average achievement level of students within a classroom (e.g., Hanushek et al., 2003; Justice et al., 2011; Mashburn et al., 2009). However, if peer effects manifest through child-to-child interaction and the transmission of specific behaviors or skills, then the average classroom achievement is a less appropriate measure of peer effect. Children are likely to have smaller social networks within the classroom consisting of proximal peers who they interact with on a regular basis. In this study, we surveyed second and third graders about who they seek help from or go to when discussing reading. The reading outcomes for these peers were identified and used to assess how they may relate to a child's own reading achievement.

A third contribution of our study is we improve on the generalizability and scope of previous research on peer effects with young children. One of the main challenges with social network analysis is the data collection procedures can be labor intensive. Respondents are required to complete network surveys about all the peers they talk to, a task that can be difficult cognitively for younger children. In this study, we collected network surveys from more than 4,000 second and third graders across nearly 300 classrooms in 41 schools and at two time points. To our knowledge, this is one of the largest studies to assess peer influence directly from the self-reports of young children. In focusing on early elementary students, the study also expands on previous studies examining peer influence in preschool (e.g., Justice et al., 2011) and upper elementary school grades (Hanushek et al., 2003).

In summary, our first research aim is to assess patterns in peer reading networks and examine whether peer reading achievement predicts children's reading achievement after controlling for prior academic achievement and individual background characteristics. Our second aim is to examine whether the influence of peers may depend on a child's individual reading ability. If struggling readers interact with peers who are stronger readers, then the contagion theory argues that these children are likely to emulate the behaviors of their peers (e.g., reading more) and become better readers over time (Harris, 2010; Hoxby & Weingarth, 2005). Research shows that learning with strong academic peers may serve as a protective factor for low-achieving achieving students (Hanushek et al., 2003; Justice et al., 2011; Mashburn et al., 2009). We asked the following three research questions:

1. To what extent do children report identifying peers who are stronger readers for help or to talk about reading?
2. To what extent does the reading achievement of peers predict the reading achievement of children, controlling for individual and classroom factors?
3. To what extent does the relationship between peer reading achievement and children's reading achievement depend on children's initial reading level?

Method

Study Context and Participants

We employed secondary data from a larger longitudinal study of an experimental reading program to reduce reading loss among low-income elementary schoolchildren (Kim et al., 2016). The initial data represent 6,383 children from 7 districts and 59 schools in North Carolina. Data collection began with 3,433 second-grade children and 2,950 third-grade children in 2013, many who were primarily from low-income (77% received free or reduced price lunch) and racial minority households (76%). About 17% spoke a non-English language at home. As part of the summer intervention, consented children in the spring were randomly assigned to receive reading lessons during the last weeks of school and 10 self-selected books in the mail each week of the summer. Books were selected based on student preference and reading level from a reading catalogue. Students in the control group participated in math lessons in the spring, but received books in the following fall. Reading assessments from the Iowa Test of Basic Skills (ITBS) were administered to all students before and after the full intervention to measure the impact on reading loss over the summer.

For the present study we used a final subsample of 4,215 total second- and third-grade students who participated in schools that used the mCLASS-DIBELS Next as a formative assessment of early reading skills, beginning in fall 2012 through spring 2013. In contrast to the ITBS that was administered to all students before and after the summer, this subgroup of students with reading scores on the mCLASS-DIBELs allowed us to assess reading achievement and peer effects during the school year. We focus on the fall to spring period, Time 1 (T1) to Time 2 (T2), to more fully capture when students are together. We note that the present study used assessment data collected before the randomized summer reading intervention, which eliminates potential spillover effects from the latter. The composition of the original and analytic subsample is comparable, with the latter having a slightly higher percentage of students on free or reduced price lunch. In the top half of Table 1, we summarize the characteristics of students in our study.

Table 1
Summary of Student and Teacher Characteristics

Characteristics	*n*	%	Mean	*SD*
Student characteristics				
Gender				
Male	2,056	49.6		
Female	2,088	50.4		
Race				
Black	1,881	45.4		
Hispanic/Latino	947	22.9		
White	612	14.8		
Other	701	16.9		
English language learner				
No	3,517	84.7		
Yes	637	15.3		
Free or reduced price lunch				
No	792	19.0		
Yes	3,379	81.0		
Grade				
Second	2,447	58.1		
Third	1,768	42.0		
Literacy skills				
Grade 2: Fall DIBELS	2,266		158.3	76.9
Grade 2: Spring DIBELS	2,341		259.8	108.8
Grade 3: Fall DIBELS	1,578		236.9	123.4
Grade 3: Spring DIBELS	1,637		334.5	131.2
Teacher characteristics				
Gender				
Female	270	94.4		
Male	16	5.6		
Race				
Black	57	19.9		
Hispanic/Latino	4	1.4		
Native American	16	5.6		
White	209	73.1		
Other	1	.4		
Age	286		38.2	11.8
Education				
Barron undergraduate ranking	251		4.0	1.1
Class size	286		14.3	4.2

Note. DIBELS = Dynamic Indicators of Basic Early Literacy Skills; Barron ranking: 1 = *most competitive*, 6 = *least competitive.*

In the bottom half of Table 1, we provide a summary of the classrooms and teachers. On average, classrooms served nearly 14 children with parental consent to participate in the study. The teacher staff was predominantly White females, comparable with most schools in the U.S. (Goldring, Gray, Bitterman, & Broughman, 2013), with an average age of 38. The teachers came from moderately competitive undergraduate schools according to the Barron ranking of universities. Because of the limited number of available classroom and teacher measures, we use the fixed effects of classrooms to control for unobserved factors that may influence children's reading achievement, in addition to children's peers.

Measures

Children's literacy skills. Our main reading outcome comes from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), an early literacy assessment similar to others like the Early Grade Reading Assessment (EGRA) commonly used in numerous countries (Dubeck & Gove, 2015). The DIBELS consists of a set of procedures assessing early literacy skills from kindergarten through sixth grade. These tests are designed to serve as one-minute fluency measures of early literacy and reading skills in the following areas: sound fluency, phoneme segmentation fluency, letter naming fluency, nonsense word fluency, oral reading fluency, and reading comprehension (Kaminski et al., 2008). Participating schools used a software version called mCLASS: DIBELS Next that provides instant analytics on students, such as progress toward benchmark goals, information for instructional lessons, and recommendations for improvement. Teachers assessed their own students in these reading areas using the mCLASS: DIBELS Next on a laptop or iPad. We used results from the fall 2012 and spring 2013 assessments as Time 1 and Time 2 measures, respectively.

Although the DIBELS provides subscale scores in the literacy areas assessed, we used a composite score that combines the different skills. The composite score (Good et al., 2011) provides a more comprehensive and reliable assessment of children's early literacy skills that is moderately correlated with standardized tests of reading comprehension (e.g., $r = .73$ between DIBELS composite and the Group Reading Assessment and Diagnostic Evaluation reading test). Depending on the grade level, the composite score may include up to six fluency areas. For the second grade assessment, the composite score consisted of measures of nonsense word fluency (i.e., basic letter-sound correspondences) and oral reading fluency (i.e., timed passage reading and retell). For the third grade assessment, the composite score included measures of oral reading fluency and reading comprehension, which entailed reading a passage and selected appropriate words for omitted text. Due to differences in the reading content covered in the DIBELS by grade level, we present the model results and analyses separately for grades two and three.

Reliability estimates (alternate-form, test-retest, and interrater) of the composite ranged from 0.88 to 0.98 across grades. Assessments of validity (content, criterion, and discriminant) with other reading assessments for separate reading components and the composite indicated that the results were at appropriate levels (see technical manual in Good et al., 2011). The DIBELS scores are summarized by both grades and time periods in Table 1.

Peer reading skills. We created a proxy for peer reading skills by first identifying the peers of each student. In the spring we administered paper and electronic surveys to children about their book preferences for the summer intervention. The survey also included a question about the children's peers that asked, "Who do you talk to about reading or, to get help, in your class?" We combined talking with someone and asking for help since research shows that the two tend to be highly correlated ($r = .87$ in Frank, Zhao, & Borman, 2004). That is, the peers that one frequently talks to also tends to be the peers that one seeks advice from. Children were asked to write the names of up to five peers in their class and instructed that it was not necessary to fill in all five spaces, a similar free-recall strategy used in the General Social Survey (Smith, McPherson, & Smith-Lovin, 2014) and Social Networks and Friendship Survey (Cairns, Cairns, Neckerman, Gest, & Gariepy, 1988). Based on the results of a pilot study using the survey questions with two classrooms, we did not include a class roster of names for children to choose from to reduce the length of the survey (Marsden, 2011). A downside of imposing a limit is it can encourage children to cite additional peers to reach the maximum or limit the true number of peers. We found some evidence

of this pattern as about 40% of children cited five peers, suggesting that some children may have needed more than five spaces whereas others may have cited additional peers to reach the limit. Unfortunately, the direction and type of measurement bias are difficult to determine in this case. We discuss these limitations in the discussion section, but our approach was a compromise between survey length and identifying children's peers. The survey itself was administered in class with a response rate of 97%.

The matching was completed using a computer database that contained the names of surveyed students for each corresponding homeroom class. All students stayed with their homeroom for literacy and regular instruction. The peer names that students listed on their survey were then matched to the names in the class roster. Sometimes there was an exact match with the name, whereas other times there were misspellings. For example, a student might spell "Jacob" as "Gacob." In cases where the misspellings were obvious, we matched the student to the peer in the database. For names that were illegible or not on the homeroom list (i.e., students may have listed peers in other classrooms), we coded the name as "unknown," which made up about 11% of all reported peers. We excluded these peers from analyses since their information, such as reading skills, could not be linked to students. About 14% of students cited no peers (i.e., 0 or unknown names), 15% for 1 matched peer, 14% for 2 peers, 14% for 3 peers, 17% for 4 peers, and 25% for 5 peers. We note that the 25% for 5 peers here differs from the 40% cited previously, which was based on 5 total peers regardless of whether the peers' names were legible and could be matched in our records. We conducted a sensitivity analysis comparing students who did not cite any peers to those who cited at least one. Results indicated no statistically significant differences between the two groups in terms of academic achievement, race, and ELL status. Students who cited no peers were less likely to be female (42% to 52%) and receiving free lunch (76% to 82%) than students who cited at least one peer.

We assumed that the peers a child identified as someone to talk about reading or acquire help from in the Spring Survey (T2) were the same peers they interacted with throughout the school year. Although peer groups are likely to change across years, these social networks are more stable when students are interacting in the same class during the year, especially for young children (Ryan & Shim, 2012). In contrast to asking children about their peers in the fall (T1) when they may not know each other well, the spring survey is likely a more accurate measure of children's peers because they would have spent nearly the entire school year together. Once the peers of each child from the Spring Survey were identified, we linked the fall DIBELS scores to each peer. Similar to other studies (Justice et al., 2011; Mashburn et al., 2009), we used the fall scores of peers as a baseline or initial measure of peer influence. The number of peers with valid scores ranged from none (9%) to five (25%) with an average of about three per child. Finally, we averaged the DIBELS scores across peers for each child to estimate the reading level of their peers, which serves as our measure of peer influence. We also explored using the highest score from peers and the score of the first peer cited as a measure of peer influence. Because these measures produced similar results (see supplemental appendix), we used the average due to its reliability. In Table 2, we present the average peer score per child by grade level.

Indegree. For each child, we calculated the number of times other children cited him or her as someone to talk about reading or seek help from. Students who have many ties, also known as indegree centrality, are considered more prominent in the network or possess specific skills that other children seek. In this study, we expected students with high indegree measures to also have stronger reading skills (Hannenman, & Riddle, 2011).

Child characteristics. We included four student characteristics in our analyses: gender, race/ethnicity (Black, Hispanic/Latino, White, and other), English language learner, and free or reduced priced lunch status. Controlling for these potential confounders can reduce the bias when estimating the relationship between peer effects and reading achievement. The background information comes from administrative records at the child's school.

Teacher and classroom characteristics. Although we were interested in estimating the influence of peers on children's reading skills, peer effects may also be confounded with teacher quality or classroom specific traits. Consequently, we included the fixed effects of classrooms in our models, which restrict our analysis of peer effects within classrooms, thereby controlling for all observed and unobserved factors (i.e., teacher and classroom characteristics)

Table 2
Summary of Student and Peer Reading Levels by Fall DIBELS Quartile and Grade

	Student		Peer			
Quartile	Mean	*SD*	Mean	*SD*	Difference	*t*
Grade 2						
1	57.5	38.1	157.2	58.5	−99.8	−32.0
2	138.6	14.4	173.0	51.5	−34.4	−14.2
3	185.7	14.0	182.7	52.6	3.0	1.2
4	254.1	35.0	193.0	54.1	61.1	21.1
Overall	158.3	76.9	176.5	55.8	−18.2	−9.1
Grade 3						
1	74.3	47.5	240.6	91.2	−166.3	−28.4
2	204.6	26.3	248.8	84.6	−44.3	−8.8
3	278.8	21.5	262.9	74.5	15.8	3.6
4	390.9	67.9	320.0	101.2	70.9	10.7
Overall	236.9	123.4	269.1	93.8	−32.3	−7.8

Note. DIBELS = Dynamic Indicators of Basic Early Literacy Skills.

that may confound the relationship between reading achievement and peers. We caution that the fixed effects and student control variables do not result in causal estimates of peer effects but help reduce selection and missing variable bias.

Analysis

To address our first research question about whether children identify peers who are stronger readers for help or to talk about reading, we conducted a descriptive analysis comparing the average achievement of children and the peers they identified. The expectation was that children would be likely to identify peers of similar or higher reading levels when seeking help about reading. We also examined the achievement level of children who were most frequently nominated by their peers as someone to seek help from or talk about reading, also known as the indegree measure (Scott & Carrington, 2011). Theoretically, children with high indegree should have higher achievement since other children are seeking them more frequently. The indegree measure is another way to assess whether children are identifying expert peers.

To address our second research question about whether peers influence children's on early reading skills, we fit the following regression model:

$$SR_{ijk} = \alpha + \beta_1 FR_{ijk} + \beta_2 FR_{ijk}^{PEER} + \gamma X_i + \delta_j + \varepsilon_{ijk} \quad (1)$$

where *SR* is the spring DIBELS reading score (T2), *FR* is the fall DIBELS reading score (T1), FR^{PEER} is the average fall DIBELS reading score (T1) of student *i*'s peers in classroom *j* of school *k*, *X* is a vector of student characteristics with γ as the parameter estimates, δ is a vector of classroom fixed effects, and ε is the error term. The model predicts the relationship between peer reading achievement in the fall and a student's spring reading achievement, while controlling for the student's own prior reading ability in the fall, student demographic characteristics, and classroom-specific effects on reading.

As an alternative strategy we considered fitting a multilevel model with students nested within classrooms and adding classroom-level covariates. However, we chose Model (1) because the classroom fixed effects allowed us to control for all observed and unobserved classroom factors that may confound the relationship between student and peer achievement. For instance, in classes where students learn better from their peers, the teachers may also be using specific instructional practices. The classroom fixed effects in δ control for these differences in teaching across classrooms. Another concern was variation in reading scores attributable to school-level differences (e.g., curriculum). However, preliminary analyses with a three-level model (students nested within classrooms within schools) indicated that less than 5% of the variation in reading scores was attributed to school differences. Of interest in Model (1) is β_2, which represents the estimated effect of peer reading achievement on student *i*'s spring reading scores. A positive parameter estimate suggests that students with peers who have high reading achievement tend to also have higher reading achievement.

To address our third research question about whether peer influence may differ for students at different levels of initial reading achievement, we supplemented Model (1) with an interaction term between student *i*'s fall reading score and student *i*'s peer fall reading score:

$$SR_{ijk} = \alpha + \beta_1 FR_{ijk} + \beta_2 FR_{ijk}^{PEER} + \beta_3 FR_{ijk} \times FR_{ijk}^{PEER} + \gamma X_i + \delta_j + \varepsilon_{ijk} \quad (2)$$

The key parameter of interest is β_3. A negative value indicates that the association between peer and children reading achievement is stronger for students with lower initial reading scores than students with higher initial reading achievement. Conversely, a positive value means that students with higher initial reading scores tend to have higher achievement when they associate with peers who are also high achievers. We conducted all analyses in Stata 14.0.

We identified missing data primarily for children's average peer reading scores (20%). Missing peer reading scores was attributable to students not completing the social network section of the survey (11%) or if we were unable to match peers to their scores (9%). To reduce potential bias from the missing data, we imputed values using chained equations in Stata 14.0 (i.e., 'mi impute chained' command) that pool together results from 20 imputed data sets (StataCorp, 2013).

Results

Research Question 1: Peer Reading Network Patterns

We begin with a descriptive summary of who children go to about reading or help. Table 2 displays children's fall DIBELS scores by quartile and the scores of the peers they identified. The results indicate that children with lower reading achievement tended to identify peers who had higher average reading scores. For instance, in the upper panel for grade two, children in the first quartile had an average score of 57.5, whereas the average of their peers' scores was nearly three times higher at 157.2, $t(1,048) = 33.2$, $p < .001$. On the other hand, we note that children in the fourth quartile with high initial reading scores did not tend to identify peers with similar or stronger achievement. Indeed, high achieving children in the fourth quartile ($M = 254.1$, $SD = 35.0$) tended to interact with peers with lower reading scores ($M = 193.0$, $SD = 54.1$). One possible explanation is that there may be fewer other high achieving students in the same classroom. Overall, in the grade two sample the average fall score for children is 158.3 whereas their peers scored an average of 176.5, or about a 0.27 *SD* difference. We see similar patterns in who children identify for help in grade three in the lower panel.

Another way to examine whether children are identifying strong readers is to focus on the achievement level of children who are in "high-demand" according to their peers. These are children who others cited as someone to go to for help with reading. The indegree measure in Table 3 indicates the number of other children who cited child *i*. For instance, about 27% of second graders had four or more peers cite them as someone they go to for help with reading. One contribution of the indegree measure is we can identify children who are experts according to their peers, which may differ from what teachers report. The results show that, across grades, children who were in higher demand tended to also have higher reading achievement than children who were identified less frequently by their peers. For instance, the average score of second graders with an indegree of 7 or more is 325.4, compared with 238.5 for children with an indegree of only 1, which indicates that children were identifying peers who were strong readers. The correlation between student indegree and DIBELS score is moderate, $r = .24$, $p < .001$.

Table 3
Summary of Indegree Frequency and Student Reading Achievement

		DIBELS Spring scores		
Indegree	Percent	Mean	*SD*	*N*
Grade 2				
0	16.0	224.4	111.6	360
1	21.9	238.5	108.6	493
2	21.2	258.0	105.2	478
3	14.0	267.3	102.3	314
4	10.8	295.6	94.3	243
5	6.3	284.7	99.9	141
6	4.0	306.1	103.1	89
7+	5.9	325.4	94.6	132
Grade 3				
0	15.3	298.8	148.2	231
1	25.6	307.6	127.7	387
2	20.5	321.8	126.4	310
3	16.4	363.6	105.5	248
4	9.7	370.6	111.2	147
5	5.4	391.9	113.5	82
6	3.0	410.4	116.6	46
7+	4.0	446.6	103.7	60

Note. DIBELS = Dynamic Indicators of Basic Early Literacy Skills.

In Figure 1, we display the social networks of two large classrooms (relative to the average classroom size of about 14 students) with the node (circle) size representing children who were strong readers based on peer nomination. We chose the two classrooms because the size permitted a better display of the network. The direction of the arrows indicates that students in each classroom identified a group of peers who others also cited as someone to go for help in reading. Children who were strong readers were also more central within the classroom. Overall, these descriptive results indicate that the children tended to identify peers who were stronger readers when they sought others to talk about reading or ask for help. Although homophily on other student traits cannot be ruled out, the descriptive results suggest that children are not necessarily affiliating with peers who have similar reading levels.

Research Question 2: Peer Effects on Reading Achievement

In Tables 4 and 5, we present results from regression models that predict the relationship between peer effect and student achievement separately for grades two and three. Each successive model includes additional covariates to reduce the bias when estimating peer effects, with classroom fixed effects controlling for unobserved differences across classrooms. Separate analyses with unconditional multilevel models justified this concern since the intraclass correlation indicates that about 18% to 24% of the variation in student scores for each grade is attributable to classroom differences. We initially identified a moderate correlation between fall peer achievement and spring student achievement, $r = .23$, $p < .001$. In Model 1, however, this relationship is no longer statistically significant when controlling for students' own prior scores in the fall ($\hat{\beta} = 0.021$, $p = .52$ 5). Results in Model 2 indicate no racial disparities in reading achievement controlling for prior achievement. Female students tended to score higher, whereas ELLs and students receiving free lunch scored lower than their counterparts. The effect of peer achievement remains statistically insignificant. We found similar results when controlling for classroom fixed effects in Model 3. The lack of a relationship between peer and student achievement is not surprising because peer scores were correlated with students' fall scores ($r = .26$). However, this suggests that the added value of peers is removed when accounting for prior achievement. We detected a similar pattern when examining peers for grade three in Table 5.

Research Question 3: Differential Peer Effects for Struggling Readers

In our third research question, we examined whether there is a stronger relationship between peer effect and student achievement for students who were struggling initial readers. In Model 4 of Table 4, the interaction term between peer achievement and initial reading achievement supports this hypothesis ($\hat{\beta} = 0.0009$, $p < .001$). The effect of peers differs by initial reading achievement, even when controlling for prior achievement and classroom fixed effects. The negative interaction indicates that children with lower initial test scores benefit more around peers with higher scores than students who were already strong readers. This finding is consistent for third graders in Table 5 ($\hat{\beta} - 0.0007, p < .001$). We display this interaction in Figure 2 for second graders. Similar to Justice et al. (2011), we define high and low initial readers as students with initial fall reading

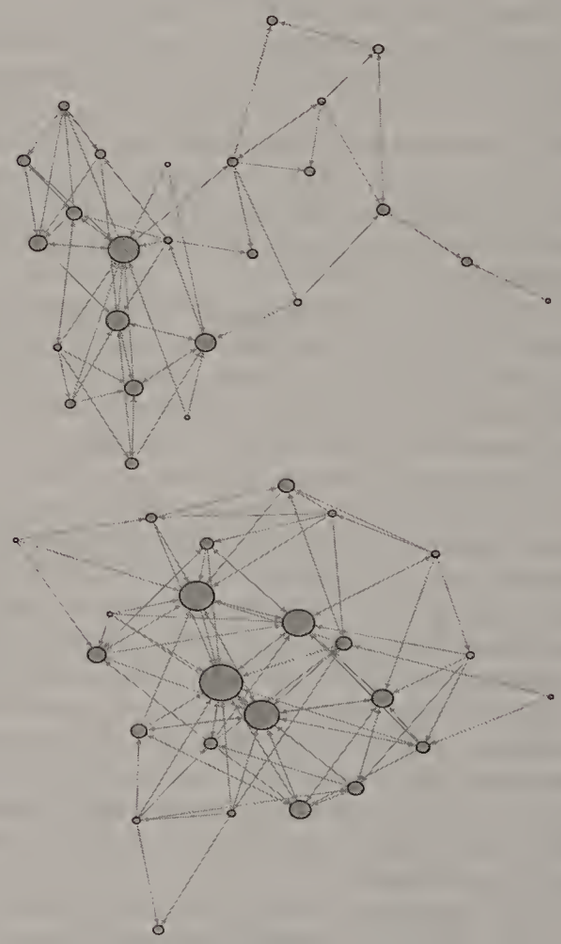


Figure 1. Examples of social networks from two classrooms. Node size and arrow direction indicate students that others report seeking about reading.

Table 4
Regression Models Predicting the Relationship Between Child Spring Literacy Skills and Peer Reading Skills in Grade 2 (n = 2,447)

Model	(1) *B* (*SE*)	(2) *B* (*SE*)	(3) *B* (*SE*)	(4) *B* (*SE*)
Peer DIBELS (Fall)	.021 (.034)	.005 (.033)	.023 (.025)	.163*** (.046)
Child DIBELS (Fall)	1.082*** (.024)	1.061*** (.026)	1.058*** (.019)	1.218*** (.047)
Female		11.144*** (2.709)	13.708*** (2.356)	13.812*** (2.343)
ELL		−11.962* (6.043)	−12.485** (4.433)	−12.003** (4.439)
Free lunch		−3.755 (4.544)	−4.486 (3.537)	−5.404 (3.523)
Hispanic/Latino		−9.034 (6.341)	4.755 (5.250)	4.483 (5.261)
Black		−2.331 (5.476)	.762 (4.297)	.525 (4.295)
Other		−11.855 (6.434)	−3.209 (5.345)	−3.173 (5.359)
Child × Peer DIBELS (Fall)				−.0009*** (.0001)
Intercept	84.646*** (7.647)	95.414*** (9.758)	86.002*** (7.172)	62.432*** (9.512)
Classroom fixed effects?	No	No	Yes	Yes

Note. Unstandardized regression coefficients with standard errors in parentheses. Models 1 and 2 use robust standard errors for student clustering at the classroom level and Models 3 and 4 use classroom fixed effects. White is the racial reference group.
* $p < .05$. ** $p < .01$. *** $p < .001$.

scores one standard deviation above and below the mean for all students, respectively. Because of a slight skew in the peer achievement variable, we used the 25th and 75th percentile to define peers of low and high reading achievement, respectively. Figure 2 shows that among students with low initial achievement in the fall, those who reported interacting with high achieving peers scored about 10 points higher on average or an effect size of about 0.09 *SD* than students who are strong readers initially, $F = 8.98$, $p = .002$. The effect size is comparable in magnitude with estimates found in similar peer influence studies (e.g., Justice et al., 2011; Mashburn et al., 2009). Students with high initial achievement perform about the same in the spring, regardless of whether they identified peers with low or strong reading skills, $F = 2.11$, $p = .148$. We see similar trends in Figure 3 for third graders, except high initial achievers scored significantly lower when around peers who are also high achievers. This is consistent with theory and research suggesting that highly skilled students may be less responsive to peer achievement (Hanushek et al., 2003).

Discussion

Because of the importance of early reading skills for later academic achievement and learning, the primary goal of this study was to examine how peers may influence students' reading skills. Although prior studies have consistently found that peers matter on a range of behavioral and academic outcomes, the theory and mechanisms through which this occurs are often overlooked. This study expanded on previous studies by directly asking a larger sample of young children across multiple classrooms and schools about the peers they talk to about reading or seek help from. Furthermore, we examined whether children reported identifying stronger readers for help and whether children experienced higher reading achievement when they identified stronger readers. A better understanding of how students report interacting and affiliating with each other is important for educators in designing practices and policies that leverage the advantages of peers. The extent to which children can identify and benefit from stronger readers has implications for teaching strategies and group activities.

Identifying Strong Readers

Our results show that across the entire sample students on average tended to report identifying peers with stronger reading achievement when asked about whom they talk to or seek help from about reading.

Table 5
Regression Models Predicting the Relationship Between Child Spring Literacy Skills and Peer Reading Skills in Grade 3 (n = 1,768)

Model	(1) *B* (*SE*)	(2) *B* (*SE*)	(3) *B* (*SE*)	(4) *B* (*SE*)
Peer DIBELS (Fall)	−.040 (.034)	−.051 (.034)	.016 (.022)	.178*** (.039)
Child DIBELS (Fall)	.915*** (.034)	.903*** (.036)	.941*** (.015)	1.126*** (.041)
Female		8.402* (3.536)	9.552** (3.243)	8.863** (3.207)
ELL		−9.151 (6.712)	−10.528 (5.878)	−10.846 (5.829)
Free lunch		−5.450 (5.203)	−8.887 (4.880)	−9.133 (4.845)
Hispanic/Latino		−19.314* (9.402)	2.209 (7.459)	.989 (7.376)
Black		−27.551*** (7.225)	−13.768* (5.426)	−15.803** (5.334)
Other		−2.266 (7.795)	3.757 (6.755)	2.454 (6.676)
Child × Peer DIBELS (Fall)				−.0007*** (.0001)
Intercept	127.325*** (12.282)	151.009*** (17.448)	115.534*** (8.949)	77.848*** (11.607)
Classroom fixed effects?	No	No	Yes	Yes

Note. Unstandardized regression coefficients with standard errors in parentheses. Models 1 and 2 use robust standard errors for student clustering at the classroom level and Models 3 and 4 use classroom fixed effects. White is the racial reference group.
* $p < .05$. ** $p < .01$. *** $p < .001$.

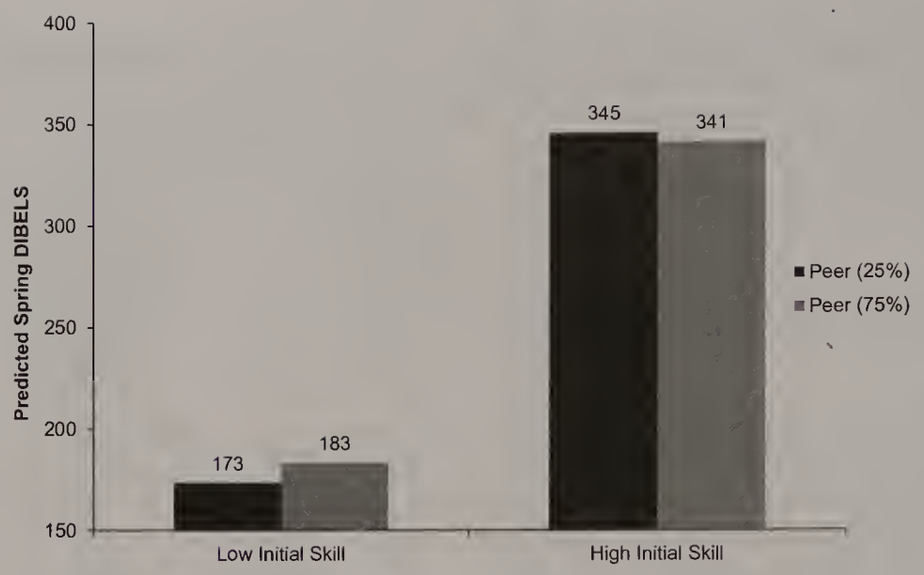


Figure 2. The interaction between children's initial (Fall) DIBELS scores (one standard deviation above and below the mean) and peers' DIBELS scores for grade two (25th and 75th percentile) when predicting children's spring DIBELs scores. Differences are significant only for low initial students ($p < .05$).

This is important because help-seeking skills in particular contribute to student motivation, learning, and later achievement, particularly in early adolescence (Karabenick & Newman, 2006). In contrast to prior studies that have shown that lower achievers are less likely to ask for help when needed (Ryan & Shim, 2012; Ryan & Shim, 2012), the present study found that students with low reading achievement tended to report interacting with peers who were stronger readers. One possible explanation for the difference in results is previous studies examined students' network behaviors based on teacher self-report, often using an overarching question about whether a student possessed appropriate help-seeking skills. The current study surveyed students directly and used social network analyses to examine their peers. Although both approaches have strengths and weaknesses, one advantage of student self-reports is they may capture peer relations within and outside the classroom that teachers may not notice. We also found that students frequently sought by peers tended to have much higher reading scores, suggesting that children in the study were effective in identifying expert readers in their class. While students on average were more likely to identify stronger readers, high achieving children tended to identify peers with lower achievement scores. For high achieving children (upper quartile), this pattern is likely attributable to having fewer high achieving peers in the same classroom. Another possibility is that these are peers they affiliate with for social reasons rather than

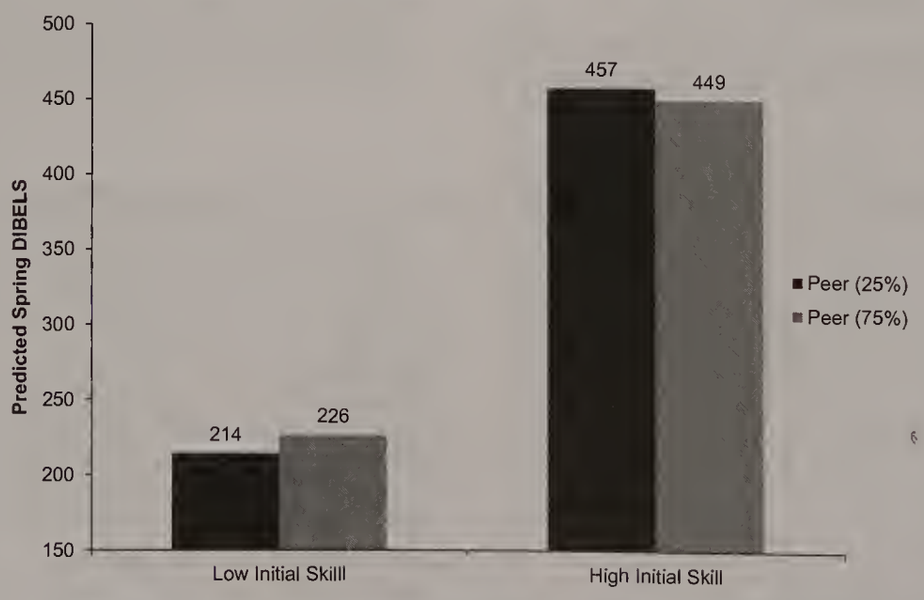


Figure 3. The interaction between children's initial (Fall) DIBELS scores (one standard deviation above and below the mean) and peers' DIBELS scores for grade three (25th and 75th percentile) when predicting children's spring DIBELs scores. Differences are significant for both groups ($p < .05$).

academic, which may confer less advantages on standardized assessments.

If children are consistently identifying peers with higher reading achievement to discuss reading or for help, particularly low-achieving children, one question for teachers is whether peer support can be more efficiently distributed. Our indegree measure in this study indicates that certain students within a classroom are frequently sought by peers. Although students may be effective in identifying peers who can help, teachers may need to provide some structure to ensure that some children are not potentially overburdened. Another implication of this finding is that teachers may want to more formally prepare children who are in "high-demand" to work with peers. This is consistent with research showing that, in addition to encouraging children to seek help or talk with peers about work, teachers often need to provide more guidance for positive feedback exchanges (e.g., Ryan & Shim, 2012). Naturally the focus has been on assisting struggling students but high-achieving peers can also assist with the process. Research on peer-mediated instruction (PMI) shows that students can be trained to effectively tutor each other or work together as partners, particularly for struggling readers across grade levels (Maheady, Harper, & Sacca, 1988; Pyle, Pyle, Lignugaris-Kraft, Duran, & Akers, 2016). Teachers can develop basic PMI routines involving verbal rehearsals of specific skills and step-by-step feedback from the tutor (Fuchs et al., 1997, 2000).

Peer Influence on Achievement

Our results indicate that peer reading achievement is positively associated with children's own achievement, but the relationship is not statistically significant after controlling for prior reading scores and background characteristics. In other words, peer effects are in part related to how strong readers tend to identify peers who also have strong reading skills or other similar characteristics and vice versa. The unique contribution of peer reading achievement is particularly difficult to detect when accounting for children's fall scores. However, we identified a small but significant interaction between peer reading achievement and children's initial reading achievement that indicated children with low reading scores appeared to benefit more from affiliating with stronger readers than children with high initial achievement. The finding that peers play a role in children's academic achievement is consistent with other recent studies examining preschoolers (Justice et al., 2011; Mashburn et al., 2009), adolescents (Calvó-Armengol, Patacchin, & Zenou, 2009), and primary school students (Hanushek et al., 2003). This pattern of peer interaction among children is also found in experimental research indicating that peer effects may manifest as novices identify experts and learn through observation (Bandura, 1977). Furthermore, peer collaboration is a context through which experts may provide direct assistance to novices (Bruner, 1975). Although our study cannot pinpoint the mechanism through which peer effects operate, our results suggest that struggling readers tend to report talking to or seeking help from expert peers, and these children appear to benefit from affiliating with high achieving peers. This is consistent with research indicating that the functional value of a peer depends on a child's perceptions of a peer's competence (e.g., Schunk, 1987). Children are more motivated to pattern their behaviors after peers who perform successfully than to emulate less competent peers.

Similar to studies that have examined peer influence for highly skilled students (Hanushek et al., 2003; Justice et al., 2011), we found that peers mattered less for high achieving readers. The reasons for why this is the case has been less explored in research. Most prior studies used the classroom average achievement as a proxy for peer achievement, which overlooks specific patterns of interactions between students. One potential reason for the null effect is that high achievers in these classrooms simply have fewer opportunities to learn from others of similar skills and, instead, may spend more time helping low achievers. The current study supports this hypothesis as high achieving readers tended to affiliate with peers with lower scores, which may have negative spillover effects whereby struggling students may pull down their scores (e.g., Fletcher, 2010). Furthermore, the high achievers were likely talking about reading with peers, as opposed to receiving help or expertise during such interactions.

Whereas many previous studies on peer influence assume that children adopt the behaviors or norms of a classroom or peer group (e.g., Justice et al., 2011; Mashburn et al., 2009; Schechter & Bye, 2007), the results in this study, based on student self-report, supports theories that peers matter because they can provide reading expertise or motivation. The findings reinforce policies and practices that aim to structure classrooms such that students have access to peers of different academic achievement levels. More importantly, the findings highlight the agency of children in interacting with peers who are strong readers. This indicates that although having access to high achieving peers matters, students are also successfully identifying them as well. Thus, classroom group or pair activities should be flexible enough to allow students to interact with peers of their choosing. Of course, raising the overall achievement of a classroom is another way to ensure that more children have access to peers who are stronger readers and can provide help. For those who are already strong readers, this also allows them more opportunities to consult with peers with similar achievement than before (as found in this study).

Implications

The findings presented here show that peer reading skills make a small but important contribution to children's reading skills, especially for those who are struggling readers. One implication is the findings support the use of peer-mediated interventions like PALS that pair students according to reading levels and provide opportunities for self-directed learning. These practices allow students to receive corrective feedback in a timely manner and engage in and respond to practice exercises (Hattie & Timperley, 2007). Policymakers and educators should also recognize the mechanisms through which peers matter to improve on classroom activities. The finding that children on average reported identifying and interacting with peers who are stronger readers does not imply that the matching of peers in structured programs like PALS is unnecessary. Instead, schools with limited resources may benefit from a mix of strategies that include providing children with classroom opportunities to engage with peers on their own (i.e., unstructured), or directly pair students who would benefit from each other.

The conclusion that struggling readers benefited more from interacting with peers of higher skill is particularly relevant for schools addressing large disparities in reading achievement. Although other targeted interventions and strategies are capable of

assisting struggling students (e.g., Edmonds et al., 2009), peers can be an efficient method that reduces the many demands placed on teachers, freeing them to focus on other instructional planning. Peer tutoring strategies have also been reported as effective across content areas and for students with disabilities or in special education settings (Klingner & Vaughn, 1996). This study suggests that a better understanding of children's peers is important for their social and academic development. High achievers, for instance, may need more opportunities to interact with similarly skilled peers to benefit academically. Although we acknowledge that teachers in this study and elsewhere are likely using these strategies to some degree, the findings provide further support of their usage when considering tradeoffs or targeting specific academic outcomes.

Teachers should also consider the importance of children's network structures and how that may benefit certain students. Children identified peers who they reported talking to or seeking help from but they are part of larger social network. According to social network theory, access to information, support, and other resources for individuals may depend on children's location within the network (Daly, 2010). That is, some students may have more favorable positions that permit easier access to high achieving students. Although the peer networks identified in this study were informal, teachers may consider ways to structure the classroom such that all children may feel more connected to each other. Indeed, the classroom network examples in Figure 1 suggest that some students may have more access to peers with stronger reader achievement while others are more isolated. Teachers should monitor these peer relationships, breaking up those that inhibit peer learning while supporting those that foster beneficial interactions to maximize opportunities for positive peer effects.

Limitations and Future Directions

There are limitations to the study that also provide several avenues for future research. First, the peer effects identified in this study are not causal. Experimental research is needed to better isolate peer effects from issues related to selection bias, and to identify the specific mediators (e.g., expert guidance, observation learning) through which peers influence the learning of individual children (Azmitia, 1988). Although we controlled for prior reading skills, student demographics, and classroom fixed effects, students who interact with strong readers are likely different in other unobserved ways that can confound the relationship between peer and student reading skills. Struggling readers who identify and interact with high-achieving peers may also have strong social skills or higher motivation that affects achievement.

Second, although the study provides insight into possible mechanisms through which peer effects may manifest, the data cannot address the type and quality of interactions among students. For instance, when students reported talking to or seeking help about reading from peers, it is unclear whether this occurred as a tutee and tutor relationship. Furthermore, while identifying peers with stronger skills is important, certain types of interactions may be more conducive to learning. Our understanding of peer interaction is also limited to student self-report, so whether children actually sought help from the peers they identified is not captured with our data. Recent studies using classroom observations of children provide a promising way to better identify peer relationships and the type of exchanges that occur between peers (Martin et al., 2013; DeLay, Hanish, Martin, & Fabes, 2016). Results from these studies using observations are consistent with this study and others using self-reported data. Delay et al. (2016), for instance, observed preschool children's peer interaction partners several times a week over one year and found that children's preschool competency was influenced by their peers' levels of competency.

A related issue is children may not necessarily be identifying peers because of their strong reading skills but peers who are more popular within the classrooms. The implication is that interacting with popular peers and adopting similar prosocial norms may benefit struggling readers. We note that the study did find that students on average tended to identify peers with stronger reading achievement. To the extent that achievement and popularity are highly correlated, which some research suggests (e.g., Meijs, Cillesssen, Scholte, Segers, & Spijerkman, 2010), this confounding effect may be attenuated and less important if students are still interacting (based on self-reports) with high achievers. Teacher observation of children's peer networks is needed to untangle whether children are identifying peers based on achievement, popularity, or other traits. Our study was also not designed to test the social contagion model, despite its importance within the peer effects literature. Future research should examine the extent to which students adopt peer or classroom norms (i.e., high expectations) and how that may influence academic achievement.

There were also limitations in our data collection and instruments. The social network survey limited students to five responses to reduce the cognitive demands for young children but future studies should consider using a roster list of students for children to choose from and identify peers. Such an approach can provide a fuller picture of children's classroom networks. Lastly, in surveying students about peers only in the spring, we had to assume that these were the same peers that students sought help from throughout the year. Our rationale for spring was students would have had a longer period to know and interact with each other, thus allowing us to examine more stable peer networks. However, future studies should survey children in the fall and spring to capture changes in relationships. The extent to which changes in peer networks with the same year and class can impact student achievement is important in deciding the level of structure and input from teachers needed to support peer activities.

Conclusion

Despite the limitations in this study, our findings provide further support that young children's reading achievement is associated with the average level of reading skills exhibited by the peers they report talking to or seeking help from, especially for struggling readers. In addition, this is one of the largest studies to directly survey young children about their reading preferences and peer networks across multiple classrooms, schools, and districts. In examining these peer network patterns, this study provides a deeper understanding of the mechanisms by which peers may influence the achievement and outcomes of children in schools. Overall, the results provide educators with a more informed view of how peer relationships form and may be leveraged within classrooms to improve learning and achievement for all children.

References

Achieve Inc. (2005). *Rising to the challenge: Are high school graduates prepared for college and work?* Washington, DC: Author.

Altermatt, E. R., & Pomerantz, E. M. (2003). The development of competence-related and motivational beliefs: An investigation of similarity and influence among friends. *Journal of Educational Psychology, 95,* 111–123. http://dx.doi.org/10.1037/0022-0663.95.1.111

Altermatt, E. R., & Pomerantz, E. M. (2005). The implications of having high-achieving versus low-achieving friends: A longitudinal analysis. *Social Development, 14,* 61–81. http://dx.doi.org/10.1111/j.1467-9507.2005.00291.x

Azmitia, M. (1988). Peer interaction and problem solving: When are two heads better than one? *Child Development, 59,* 87–96. http://dx.doi.org/10.2307/1130391

Bandura, A. (1977). *Social learning theory.* Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Bruner, J. S. (1975). From communication to language: A psychological perspective. *Cognition, 3,* 255–287. http://dx.doi.org/10.1016/0010-0277(74)90012-2

Cairns, R. B., Cairns, B. D., Neckerman, H. J., Gest, S. D., & Gariépy, J.-L. (1988). Social networks and aggressive behavior: Peer support or peer rejection? *Developmental Psychology, 24,* 815–823. http://dx.doi.org/10.1037/0012-1649.24.6.815

Calvó-Armengol, A., Patacchini, E., & Zenou, Y. (2009). Peer effects and social networks in education. *The Review of Economic Studies, 76,* 1239–1267. http://dx.doi.org/10.1111/j.1467-937X.2009.00550.x

Christakis, N. A., & Fowler, J. H. (2013). Social contagion theory: Examining dynamic social networks and human behavior. *Statistics in Medicine, 32,* 556–577. http://dx.doi.org/10.1002/sim.5408

Daly, A. J. (2010). *Social network theory and educational change.* Cambridge, MA: Harvard Education Press.

DeLay, D., Hanish, L. D., Martin, C. L., & Fabes, R. A. (2016). Peer effects on Head Start children's preschool competency. *Developmental Psychology, 52,* 58–70. http://dx.doi.org/10.1037/dev0000066

Dubeck, M. M., & Gove, A. (2015). The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations. *International Journal of Educational Development, 40,* 315–322. http://dx.doi.org/10.1016/j.ijedudev.2014.11.004

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C., Cable, A., Tackett, K. K., & Schnakenberg, J. W. (2009). A synthesis of reading interventions and effects on reading comprehension outcomes for older struggling readers. *Review of Educational Research, 79,* 262–300. http://dx.doi.org/10.3102/0034654308325998

Ellis, S., & Rogoff, B. (1986). Problem solving in children's management of instruction. In E. Mueller & C. Cooper (Eds.), *Process and outcome in peer relationships* (pp. 301–326). New York, NY: Academic Press.

Entwisle, D. R., & Alexander, K. L. (1994). Winter setback: The racial composition of schools and learning to read. *American Sociological Review, 59,* 446–460. http://dx.doi.org/10.2307/2095943

Fletcher, J. (2010). Spillover effects of inclusion of classmates with emotional problems on test scores in early elementary school. *Journal of Policy Analysis and Management, 29,* 69–83. http://dx.doi.org/10.1002/pam.20479

Frank, K. A., Muller, C., Schiller, K. S., Riegle-Crumb, C., Mueller, A. S., Crosnoe, R., & Pearson, J. (2008). The social dynamics of mathematics coursetaking in high school. *American Journal of Sociology, 113,* 1645–1696. http://dx.doi.org/10.1086/587153

Frank, K. A., Zhao, Y., & Borman, K. (2004). Social capital and the diffusion of innovations within organizations: The case of computer technology in schools. *Sociology of Education, 77,* 148–171. http://dx.doi.org/10.1177/003804070407700203

Fuchs, D., Fuchs, L. S., & Burish, P. (2000). Peer-assisted learning strategies: An evidence-based practice to promote reading achievement. *Learning Disabilities Research & Practice, 15,* 85–91. http://dx.doi.org/10.1207/SLDRP1502_4

Fuchs, D., Fuchs, L. S., Mathes, P. G., & Simmons, D. C. (1997). Peer-assisted learning strategies: Making classrooms more responsive to diversity. *American Educational Research Journal, 34,* 174–206. http://dx.doi.org/10.3102/00028312034001174

Fuchs, D., Fuchs, L. S., Thompson, A., Svenson, E., Yen, L., Al Otaiba, S., . . . Saenz, L. (2000). Peer-assisted learning strategies in reading: Extensions for kindergarten, first grade, and high school. *Remedial and Special Education, 22,* 15–21. http://dx.doi.org/10.1177/07419325010220010301

Fujimoto, K., Unger, J. B., & Valente, T. W. (2012). A network method of measuring affiliation-based peer influence: Assessing the influences of teammates' smoking on adolescent smoking. *Child Development, 83,* 442–451.

Goldring, R., & Gray, L., & Bitterman, A. (2013). *Characteristics of public and private elementary and secondary school teachers in the United States: Results from the Schools and Staffing Survey.* Washington, DC: National Center for Education Statistics.

Good, R. H., Kaminski, R. A., Cummings, E., Dufour-Martel, C., Petersen, K., Powell-Smith, K., . . . Wallin, J. (2011). *DIBELS Next assessment manual.* Eugene, OR: Dynamic Measurement Group.

Greenwood, C. R., Delquadri, J. C., & Hall, R. V. (1989). Longitudinal effects of classwide peer tutoring. *Journal of Educational Psychology, 81,* 371–383. http://dx.doi.org/10.1037/0022-0663.81.3.371

Hanish, L. D., Martin, C. L., Fabes, R. A., Leonard, S., & Herzog, M. (2005). Exposure to externalizing peers in early childhood: Homophily and peer contagion processes. *Journal of Abnormal Child Psychology, 33,* 267–281. http://dx.doi.org/10.1007/s10802-005-3564-6

Hannenman, R. A., & Riddle, M. (2011). Concepts and measures of basic network analysis. In J. Scott & P. Carrington (Eds.), *The Sage handbook of social network analysis* (pp. 341–369). Thousand Oaks, CA: Sage.

Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Psychometrics, 18,* 527–544.

Harris, D. N. (2010). How do school peers influence student educational outcomes? Theory and evidence from economics and other social sciences. *Teachers College Record, 112,* 1163–1197.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77,* 81–112.

Henry, G. T., & Rickman, D. K. (2007). Do peers' influence children's skill development in preschool? *Economics of Education Review, 26,* 100–112. http://dx.doi.org/10.1016/j.econedurev.2005.09.006

Hong, G., Corter, C., Hong, Y., & Pelletier, J. (2012). Differential effects of literacy instruction time and homogenous ability group in kindergarten classroom: Who will benefit? Who will suffer? *Educational Evaluation and Policy Analysis, 34,* 69–88. http://dx.doi.org/10.3102/0162373711424206

Hoxby, C. M., & Weingarth, G. (2005). *Taking race out of the equation: School reassignment and the structure of peer effects.* Unpublished manuscript.

Jencks, C., & Mayer, S. E. (1990). The social consequences of growing up in a poor neighborhood. In L. E. Lynn, Jr. & M. G. H. McGeary (Eds.), *Inner city poverty in the United States* (pp. 111–186). Washington, DC: National Academy Press.

Justice, L. M., Logan, J. A. R., Lin, T.-J., & Kaderavek, J. N. (2014). Peer effects in early childhood education: Testing the assumptions of special-education inclusion. *Psychological Science, 25,* 1722–1729. http://dx.doi.org/10.1177/0956797614538978

Justice, L. M., Petscher, Y., Schatschneider, C., & Mashburn, A. (2011). Peer effects in preschool classrooms: Is children's language growth

associated with their classmates' skills? *Child Development, 82,* 1768–1777. http://dx.doi.org/10.1111/j.1467-8624.2011.01665.x

Kamil, M. L. (2003). *Adolescents and literacy: Reading for the 21st century*. Washington, DC: Alliance for Excellent Education.

Kaminski, R., Cummings, K. D., Powell-Smith, K. A., & Good, R. H. (2008). Best practices in using dynamic indicators of basic early literacy skills for formative assessment and evaluation. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology* (Vol. V, pp. 1181–1203). Bethesda, MD: National Association of School Psychologists.

Kandel, D. B. (1978). Homophily, selection, and socialization in adolescent friendships. *American Journal of Sociology, 84,* 427–436. http://dx.doi.org/10.1086/226792

Karabenick, S. A., & Newman, R. S. (Eds.). (2006). *Help seeking in academic settings: Goals, groups, and contexts*. Mahwah, NJ: Erlbaum.

Kim, J. S., Guryan, J., White, T. G., Quinn, D. M., Capotosto, L., & Kingston, H. C. (2016). Delayed effects of a low-cost and large-scale summer reading intervention on elementary school children's reading comprehension. *Journal of Research on Educational Effectiveness, 9,* 1–22. http://dx.doi.org/10.1080/19345747.2016.1164780

Kimelberg, S. D., & Billingham, C. M. (2013). Attitudes toward diversity and the school choice process: Middle-class parents in a segregated urban public school district. *Urban Education, 48,* 198–231. http://dx.doi.org/10.1177/0042085912449629

Kindermann, T. A. (2007). Effects of naturally existing peer groups on changes in academic engagement in a cohort of sixth graders. *Child Development, 78,* 1186–1203. http://dx.doi.org/10.1111/j.1467-8624.2007.01060.x

Klingner, J. K., & Vaughn, S. (1996). Reciprocal teaching of reading comprehension strategies for students with disabilities who use English as a second language. *Elementary School Journal, 96,* 275–293.

Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in educational research. *Educational Researcher, 43,* 242–252. http://dx.doi.org/10.3102/0013189X14539189

Maheady, L., Harper, G. F., & Sacca, M. K. (1988). Peer-mediated instruction: A promising approach to meeting the diverse needs of LD adolescents. *Learning Disability Quarterly, 11,* 108–113. http://dx.doi.org/10.2307/1510988

Marsden, P. V. (2011). Survey methods for network data. In J. Scott & P. Carrington (Eds.), *The Sage handbook of social network analysis* (pp. 370–388). Thousand Oaks, CA: Sage.

Martin, C. L., Kornienko, O., Schaefer, D. R., Hanish, L. D., Fabes, R. A., & Goble, P. (2013). The role of sex of peers and gender-typed activities in young children's peer affiliative networks: A longitudinal analysis of selection and influence. *Child Development, 84,* 921–937. http://dx.doi.org/10.1111/cdev.12032

Mashburn, A. J., Justice, L. M., Downer, J. T., & Pianta, R. C. (2009). Peer effects on children's language achievement during pre-kindergarten. *Child Development, 80,* 686–702. http://dx.doi.org/10.1111/j.1467-8624.2009.01291.x

Meijs, N., Cillessen, A. H. N., Scholte, R. H. J., Segers, E., & Spijkerman, R. (2010). Social intelligence and academic achievement as predictors of adolescent popularity. *Journal of Youth and Adolescence, 39,* 62–72. http://dx.doi.org/10.1007/s10964-008-9373-9

National Center for Education Statistics. (2013). *The national's report card: A first look 2013 mathematics and reading* (NCES 2014–451). Washington, DC: Institute of Education Sciences, U.S. Department of Education.

Pyle, D., Pyle, N., Lignugaris-Kraft, B., Duran, L., & Akers, J. (2016). Academic effects of peer-mediated interventions with English language learners: A research synthesis. *Review of Educational Research*. Advance online publication. http://dx.doi.org/10.3102/0034654316653663

Roda, A., & Wells, A. S. (2013). School choice policies and racial segregation: Where white parents' good intentions, anxiety, and privilege collide. *American Journal of Education, 119,* 261–293. http://dx.doi.org/10.1086/668753

Ryan, A. M. (2001). The peer group as a context for the development of young adolescent motivation and achievement. *Child Development, 72,* 1135–1150. http://dx.doi.org/10.1111/1467-8624.00338

Ryan, A. M., & Shim, S. S. (2012). Changes in help seeking from peers during early adolescence: Associations with changes in achievement and perceptions of teachers. *Journal of Educational Psychology, 104,* 1122–1134. http://dx.doi.org/10.1037/a0027696

Sacerdote, B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *The Quarterly Journal of Economics, 116,* 681–704. http://dx.doi.org/10.1162/00335530151144131

Schechter, C., & Bye, B. (2007). Preliminary evidence for the impact of mixed-income preschools on low-income children's language growth. *Early Childhood Research Quarterly, 22,* 137–146. http://dx.doi.org/10.1016/j.ecresq.2006.11.005

Schunk, D. H. (1987). Peer models and children's behavioral change. *Review of Educational Research, 57,* 149–174. http://dx.doi.org/10.3102/00346543057002149

Schunk, D. H. (1998). Teaching elementary students to self-regulate practice of mathematical skills with modeling. In D. H. Schunk & B. J. Zimmerman (Eds.), *Self-regulated learning, from teaching to self-reflective practice* (pp. 137–159). New York, NY: Guilford Press.

Schunk, D. H., & Hanson, A. R. (1989). Self-modeling and children's cognitive skill learning. *Journal of Educational Psychology, 81,* 155–163. http://dx.doi.org/10.1037/0022-0663.81.2.155

Scott, J., & Carrington, P. J. (2011). *The Sage handbook of social network analysis*. Thousand Oaks, CA: Sage.

Smith, J. A., McPherson, M., & Smith-Lovin, L. (2014). Social distance in the United States: Sex, race, religion, and education homophily among confidants, 1985–2004. *American Sociological Review, 79,* 432–456. http://dx.doi.org/10.1177/0003122414531776 .

Snow, C. E., & Biancarosa, G. (2003). *Adolescent literacy and the achievement gap: What do we know and where do we go from here?* New York, NY: Carnegie Corporation of New York.

StataCorp. (2013). *Stata: Release 13* [Statistical software]. College Station, TX: StataCorp LP.

Vu, J. A., & Locke, J. J. (2014). Social network profiles of children in early elementary school classrooms. *Journal of Research in Childhood Education, 28,* 69–84. http://dx.doi.org/10.1080/02568543.2013.850128

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Wood, D. (1980). Teaching the young child: Some relationships between social interaction, language, and thought. In D. Olson (Ed.), *The social foundations of language and thought* (pp. 280–298). New York, NY: Norton.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology & Psychiatry & Allied Disciplines, 17,* 89–100. http://dx.doi.org/10.1111/j.1469-7610.1976.tb00381.x

Zimmer, R. W., & Toma, E. F. (2000). Peer effects in private and public schools across countries. *Journal of Policy Analysis and Management, 19,* 75–92. http://dx.doi.org/10.1002/(SICI)1520-6688(200024)19:1<75::AID-PAM5>3.0.CO;2-W

Received April 12, 2016
Revision received October 25, 2016
Accepted October 28, 2016 ■

Call for Papers

A focused collection of qualitative studies in the psychological sciences: Reasoning and participation in formal and informal learning environments

Journal of Educational Psychology
Guest Editors: Tanner LeBaron Wallace and Eric Kuo

Reasoning and participation are two central topics of education research in the psychological sciences. Understanding the mechanisms that govern thought and reasoning has long been a core enterprise of educational psychology and, over time, more modern views on learning have promoted participation as a key feature for research—either as a facilitator of learning, a practice to be learned, or as an operationalization of learning itself.

We are pleased to announce a focused collection highlighting qualitative studies of reasoning and participation in formal and informal learning environments. By inviting studies incorporating qualitative methods, we aim to complement the experimental and longitudinal statistical research on these topics that is typically published in this journal. We encourage submission of papers focused on the following (or closely related) topics:

- Student reasoning and/or participation in novel learning environments or activities
- The relations between student reasoning, motivation, identity, and participation
- Student perceptions and meaning-making during participatory experiences
- Dynamic models of student reasoning that are grounded in data
- Explanatory accounts for how and why participation is successful (or not)
- Identifying new goals or targeted outcomes for reasoning or participation

We especially welcome qualitative studies that demonstrate the possibilities for unique discovery afforded by inductive analysis of rich data sources (e.g., real-time recordings of student reasoning, participation, discourse, and physical action, students' meaning-making anchored to particular interactions experienced). This collection will highlight the benefits of qualitative methods for extending and deepening theoretical and empirical understandings of reasoning and participation in both formal and informal learning environments.

The deadline for manuscript submissions is **March 1, 2018**. We invite authors to contact the Guest Editors of this collection, Tanner LeBaron Wallace (twallace@pitt.edu) and Eric Kuo (erickuo@pitt.edu), for discussion on how to maximize alignment between their submissions and this focused collection, though it is not required. Please follow both APA guidelines as well as specific submission criteria for the journal. When submitting manuscripts, please also indicate your intent to submit to this focused collection in the required cover letter.

All manuscripts must be submitted electronically at http://www.editorialmanager.com/edu. In the submission portal, please select the article type "Special Section: Reasoning & Participation – Qualitative." For more information on the *Journal of Educational Psychology*, please visit http://www.apa.org/pubs/journals/edu/.

Instructions to Authors
Journal of Educational Psychology
www.apa.org/pubs/journals/edu

The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

Manuscript preparation. Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (6th ed.). Manuscripts may be copyedited for bias-free language (see pp. 70–77 of the *Publication Manual*). Formatting instructions (all copy must be double-spaced) and instructions on the preparation of tables, figures, references, metrics, and abstracts appear in the *Manual*. For APA's Checklist for Manuscript Submission, see www.apa.org/pubs/journals/edu. **Abstract and keywords.** All manuscripts must include an abstract containing a maximum of 250 words typed on a separate page. After the abstract, please supply up to five keywords or brief phrases. **References.** References should be listed in alphabetical order. Each listed reference should be cited in text, and each text citation should be listed in the References. Basic formats are as follows:

Hughes, G., Desantis, A., & Waszak, F. (2013). Mechanisms of intentional binding and sensory attenuation: The role of temporal prediction, temporal control, identity prediction, and motor prediction. *Psychologial Bulletin, 139,* 133–151. http://dx.doi.org/10.1037/a0028566

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.

Gill, M. J., & Sypher, B. D. (2009). Workplace incivility and organizational trust. In P. Lutgen-Sandvik & B. D. Sypher (Eds.), *Destructive organizational communication: Processes, consequences, and constructive ways of organizing* (pp. 53–73). New York, NY: Taylor & Francis.

Adequate description of participants is critical to the science and practice of educational psychology; this allows readers to assess the results, determine generalizability of findings, and make comparisons in replications, extensions, literature reviews, or secondary data analyses. Authors should see guidelines for sample–subject description in the *Manual*. Appropriate indexes of effect size or strength of relationship should be incorporated in the results section of the manuscript (see p. 34 of the *Manual*). Information that allows the reader to assess not only the significance but also the magnitude of the observed effects or relationships clarifies the importance of the findings. **Figures.** Graphics files are welcome if supplied in TIFF or EPS format. APA's policy on publication of color figures is available at http://www.apa.org/pubs/authors/instructions.aspx?item=6.

Publication policies. APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. APA policy regarding posting articles on the Internet may be found at www.apa.org/pubs/authors/posting.aspx. In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 8.13). As this is a primary journal that publishes original material only, APA policy prohibits publication of any manuscript or data that have already been published in whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data on which their article depends. In addition, APA Ethical Principles specify that "after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 8.14). Authors must have available their data throughout the editorial review process and for at least 5 years after the date of publication.

Masked review policy. The *Journal* has a masked review policy, which means that the identities of both authors and reviewers are masked. Every effort should be made by the authors to see that the manuscript itself contains no clues to their identities. Authors should never use first person (*I, my, we, our*) when referring to a study conducted by the author(s) or when doing so reveals the authors' identities, e.g., "in our previous work, Johnson et al., 1998 reported that . . .." Instead, references to the authors' work should be in third person, e.g., "Johnson et al. (1998) reported that . . .." The authors' institutional affiliations should also be masked in the manuscript. Authors submitting manuscripts are required to include in the cover letter the title of the manuscript along with all authors' names and institutional affiliations. However, the first page of the manuscript should omit the authors' names and affiliations, but should include the title of the manuscript and the date it is submitted. Responsibility for masking the manuscript rests with the authors; manuscripts will be returned to the author if not appropriately masked. If the manuscript is accepted, authors will be asked to make changes in wording so that the paper is no longer masked. Authors are required to state in writing that they have complied with APA ethical standards in the treatment of their sample, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained at www.apa.org/ethics/ or by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242. APA requires authors to reveal any possible conflict of interest in the conduct and reporting of research (e.g., financial interests in a test procedure, funding by pharmaceutical companies for drug research). Authors of accepted manuscripts will be required to transfer copyright to APA.

Permissions. Authors of accepted papers must obtain and provide to the editor on final acceptance all necessary permissions to reproduce in print and electronic form any copyrighted work, including test materials (or portions thereof), photographs, and other graphic images (including those used as stimuli in experiments). On advice of counsel, APA may decline to publish any image whose copyright status is unknown.

Supplemental materials. APA can place supplementary materials online, which will be available via the published article in the PsycARTICLES database. To submit such materials, please see www.apa.org/pubs/authors/supp-material.aspx for details. Authors of accepted papers will be asked to work with the editor and production staff to provide supplementary materials as appropriate.

Submission. Authors should submit their manuscripts electronically via the Manuscript Submission Portal at www.apa.org/pubs/journals/edu/index.aspx (follow the link for submission under Instructions to Authors). General correspondence may be addressed to the incoming editorial office at AConley@apa.org.

transforming the future

AMERICAN PSYCHOLOGICAL FOUNDATION

"We would like to thank the APF and Division 39 for its generous support of our work. With these funds, we were able to expand our work with a vulnerable and underserved population."

THE KEDZIE CENTER
2016 APF DIVISION 39 GRANT RECIPIENT

ABOUT THE AMERICAN PSYCHOLOGICAL FOUNDATION

The American Psychological Foundation (APF) provides financial support for innovative research and programs for students and early career psychologists working to make a difference in people's lives. APF grantees work on issues of pivotal concern, including preventing violence, helping children, fostering the connection between behavior and health, fighting stigma and prejudice, and helping with the long-term effects of disaster. An APF grant can unlock discoveries, lead to federal funding, and help solve some of society's thorniest problems. APF's work would not be possible without the generosity of psychologists from around the world.

For more information about APF programs and how you can support the future of psychology, please call (202) 336-5843 or visit **www.apa.org/apf.**

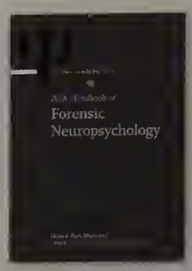
NEW RELEASES

from the American Psychological Association

APA Handbook of Forensic Neuropsychology

Editor-in-Chief Shane S. Bush
2018. 528 pages. Hardcover.

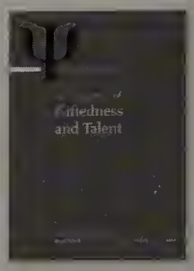
Series: *APA Handbooks in Psychology®*
List: $199.00 | APA Member/Affiliate: $129.00
ISBN 978-1-4338-2694-8 | Item # 4311532



APA Handbook of Giftedness and Talent

Editor-in-Chief Steven I. Pfeiffer
2018. 704 pages. Hardcover.

Series: *APA Handbooks in Psychology®*
List: $199.00 | APA Member/Affiliate: $129.00
ISBN 978-1-4338-2696-2 | Item # 4311533

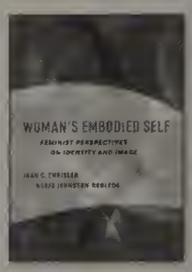


Woman's Embodied Self

Feminist Perspectives on Identity and Image

Joan C. Chrisler and Ingrid Johnston-Robledo
2018. 367 pages. Hardcover.

Series: *Psychology of Women*
List: $89.95 | APA Member/Affiliate: $44.95
ISBN 978-1-4338-2712-9 | Item # 4318148



Mindful Sport Performance Enhancement

Mental Training for Athletes and Coaches

Keith A. Kaufman, Carol R. Glass, and Timothy R. Pineau
2018. 431 pages. Hardcover.

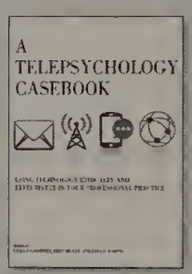
List: $89.95 | APA Member/Affiliate: $49.95
ISBN 978-1-4338-2787-7 | Item # 4317456
Available on Amazon Kindle®

A Telepsychology Casebook

Using Technology Ethically and Effectively in Your Professional Practice

Linda F. Campbell, Fred Millán, and Jana N. Martin
2018. 289 pages. Paperback.

List: $59.95 | APA Member/Affiliate: $39.95
ISBN 978-1-4338-2706-8 | Item # 4317443



Designing and Proposing Your Research Project

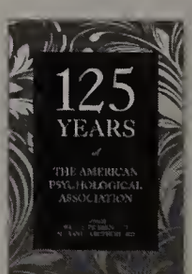
Jennifer Brown Urban and Bradley Matheus Van Eeden-Moorefield
2018. 146 pages. Paperback.

Series: *Concise Guides to Conducting Behavioral, Health, and Social Science Research*
List: $29.95 | APA Member/Affiliate: $25.95
ISBN 978-1-4338-2708-2 | Item # 4313045

125 Years of the American Psychological Association

Edited by Wade E. Pickren and Alexandra Rutherford
2018. 625 pages. Hardcover.

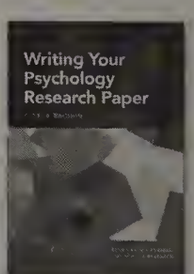
List: $69.95 | APA Member/Affiliate: $34.95
ISBN 978-1-4338-2791-4 | Item # 4316182
Available on Amazon Kindle®



Writing Your Psychology Research Paper

Scott A. Baldwin
2018. 126 pages. Paperback.

Series: *Concise Guides to Conducting Behavioral, Health, and Social Science Research*
List: $29.95 | APA Member/Affiliate: $25.95
ISBN 978-1-4338-2707-5 | Item # 4313044



TO ORDER: 800-374-2721 • www.apa.org/pubs/books

AMERICAN PSYCHOLOGICAL ASSOCIATION | 1892 2017

AD3161